PB90-225764

ASA/EPA Conferences on Interpretation of Environmental Data: IV. Compliance Sampling, October 5-6, 1987

(U.S.) Environmental Protection Agency, Washington, DC

1987

U.S. Department of Commerce
National Technical Information Service
NTIS

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)
PB90 225764/AS

2. REPORT DATE
1987

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
ASA/EPA Conferences on Interpretation of Environmental Data
IV compliance sampling Oct 5-6th, 1987

5. FUNDING NUMBERS

6. AUTHOR(S)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Office of Policy, Planning and Evaluation
EPA
401 M St, SW, Wash. D.C. 20460

8. PERFORMING ORGANIZATION REPORT NUMBER
EPA-230-03-047

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Same as in item 7

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Release unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)
from iii

The general theme of the papers and associated discussions is the design and interpretation of environmental regulations that incorporate, from the outset, statistically valid compliance verification procedures. Statistical aspects of associated compliance monitoring programs are considered. Collectively the papers deal with a wide variety of environmental concerns including various novel approaches to air emissions regulations and monitoring, spatial sampling of soil, incorporation of potential health effects considerations into the design of monitoring programs, and considerations in the statistical evaluation of analytical laboratory performance.

14. SUBJECT TERMS

15. NUMBER OF PAGES
172

16. PRICE CODE

17. SECURITY CLASSIFICATION
UNC

18. SECURITY CLASSIFICATION
UNC

19. SECURITY CLASSIFICATION
[illegible]

20. LIMITATION OF ABSTRACT
UL

PB90-225764

United States
Environmental Protection
Agency

Office of Policy, Planning
and Evaluation
Washington, DC 20460

EPA-230-03-047

Statistical Policy Branch

EPA

ASA/EPA Conferences on Interpretation of Environmental Data

IV Compliance Sampling
October 5 -6th, 1987

STA
TIST
ICS
1987

REPRODUCED BY
U.S. DEPARTMENT OF COMMERCE
NATIONAL TECHNICAL
INFORMATION SERVICE
SPRINGFIELD, VA 22161

DISCLAIMER

This document has not undergone final review within EPA and should not be used to infer EPA approval of the views expressed.

United States
Environmental Protection
Agency

Office of Policy, Planning
and Evaluation
Washington, DC 20460

EPA-230-03-047

Statistical Policy Branch

EPA

ASA/EPA Conferences on Interpretation of Environmental Data

IV Compliance Sampling
October 5 -6th, 1987

STA
TIST
ICS
1987

PREFACE

This volume is a compendium of the papers and commentaries that were presented at the fourth in a series of conferences on interpretation of environmental data conducted by the American Statistical Association and the U.S. Environmental Protection Agency's Statistical Policy Branch of the Office of Standards and Regulations/Office of Policy, Planning, and Evaluation. The ASA Committee on Statistics and the Environment developed this series and has general responsibility for it.

The purpose of these conferences is to provide a forum in which professionals from the academic, private, and public sectors exchange ideas on statistical problems that confront EPA in its charge to protect the public and the environment through regulation of toxic exposures. They provide a unique opportunity for Agency statisticians and scientists to interact with their counterparts in the private sector.

BDS

The eight papers and accompanying discussions in this volume of proceedings are about "compliance sampling" to determine how well environmental standards are met. These papers provide valuable guidance in the planning of future environmental studies. The papers address many aspects of compliance, and are intended for statisticians involved in planning how to ascertain general levels of compliance and identify noncompliers for special attention. Such work is inherently statistical and must be based on anticipation of the statistical analysis to be performed so that the necessary data can be collected. These proceedings should help the statistician anticipate the analyses to be performed. In addition, the papers discuss implications for new studies. No general prescriptions are offered; none may be possible.

The emphases in these papers are quite different. No two authors have chosen the same aspect of compliance to examine. This diversity suggests that a major challenge is to consider carefully each study aspect in the planning process. Meeting this challenge will require a high degree of professionalism from the statistical community.

The conference itself and these proceedings are primarily the result of the efforts of the authors and discussants. The discussants not only describe how their views differ from those of the authors, but provided independent ideas as well. The coordination of the conference and of the publication of the proceedings was carried out by Mary Esther Barnes and Lee L. Decker of the ASA staff.

The views presented in this conference are those of individual writers and should not be construed as reflecting the official position of any agency or organization.

This fourth conference, "Compliance Sampling," was held in October 1987. Others were the first conference, "Current Assessment of Combined Toxicant Effects," in May 1986, the second , "Statistical Issues in Combining Environmental Studies," in October 1986, and the third , "Sampling and Site Selection in Environmental Studies," in May 1987.

John C. Bailar III, Editor
Chair, ASA Committee on Statistics and the Environment
Department of Epidemiology and Biostatistics, McGill University
and
Office of Disease Prevention and Health Promotion
U.S. Department of Health and Human Services

INTRODUCTION

The general theme of the papers and associated discussions is the design and interpretation of environmental regulations that incorporate, from the outset, statistically valid compliance verification procedures. Statistical aspects of associated compliance monitoring programs are considered. Collectively the papers deal with a wide variety of environmental concerns including various novel approaches to air emissions regulations and monitoring, spatial sampling of soil, incorporation of potential health effects considerations into the design of monitoring programs, and considerations in the statistical evaluation of analytical laboratory performance.

Several papers consider aspects of determining appropriate sampling frequencies. Allan Marcus discusses how response time frames of potential biological and health effects due to exposures may be used to decide upon appropriate monitoring interval time frames. He demonstrates how biokinetic modeling may be used in this regard.

Neil Frank and Tom Curran discuss factors influencing required sampling frequencies to detect particulate levels in air. They emphasize the need to specify compliance monitoring requirements right at the time that the air quality standard is being formulated. They suggest an adaptive monitoring approach based on site specific requirements. Those sites that are clearly well above or well below the standard need be sampled relatively infrequently. Those sites that straddle the standard should be sampled more frequently to decrease the probabilities of misclassification of attainment/nonattainment status.

Tom Hammerstrom and Ron Wyzga discuss strategies to accommodate situations when Allan Marcus' recommendations for determining sampling frequency have not been followed, namely when monitoring data averaging time intervals are very long relative to exposure periods that may result in adverse physiological and health consequences. For example, air monitoring data may be averaged over one hour intervals but respiratory symptoms may be related to the highest five minutes of exposure during that hour. The authors model the relationships between peak five minute average concentration during an hour and the overall one hour average concentration under various stochastic process assumptions. They combine monitoring and modeling to predict short term peak concentrations on the basis of observed longer term average concentrations.

Bill Nelson discusses statistical aspects of personal monitoring and monitoring "microenvironments" such as homes and workplaces to assess total personal exposure. Such data are very useful for the exposure assessment portions of risk assessment. Dr. Nelson compares and contrasts personal monitoring with the more traditional area monitoring. The availability of good personal exposure data would permit much greater use of human epidemiologic data in place of animal toxicologic data in risk assessment.

Richard Gilbert, M. Miller, and H. Meyer discuss statistical aspects of sampling "frequency" determination in the spatial sense. They consider the development of a soil sampling program to estimate levels of radioactive solid contamination. They discuss the use of multilevel acceptance sampling plans to determine the compliance status of individual soil plots. These plans have sufficient sensitivity to distinguish between compliant/noncompliant plots yet result in substantial sample size economies relative to more naive single stage plans.

John Holley and Barry Nussbaum present an economist's approach to environmental regulation. The "bubble" concept specifies that average environmental standards must be maintained across a dimension such as area, time, auto fleet, or industry group. This dimension constitutes the "bubble." Lack of compliance in one part of the bubble may be offset by greater than minimum compliance in other parts. Emissions producers have the option to trade, sell or purchase emissions "credits" with, from, or to other emissions producers in the bubble. Alternatively, they may "bank" emissions "credits" for use in a future time period. Such an approach to regulation greatly enhances the emissions producers' flexibility, as a group, to configure their resources so as to most economically comply with the overall standard.

Soren Bisgaard and William Hunter discuss statistical aspects of the formulation of environmental regulations. They emphasize that the regulations, including their associated compliance monitoring requirements, should be designed to have satisfactory statistical characteristics. One approach to this is to design regulations that have operating characteristic curves of desired shape. Alternative candidate formulations can be compared in terms of the shapes of their associated operating characteristic curves.

Bert Price discusses yet another statistical aspect of environmental regulation; evaluating the capabilities of analytical laboratories. He contrasts and compares strategies to evaluate individual laboratories based only on their own bias and variability characteristics (intralaboratory testing) with strategies that evaluate laboratories as a group (interlaboratory testing). Price's paper has commonality with that of Bisgaard and Hunter in that he argues that first the operating characteristic of a regulation needs to be specified. This specification is then used to determine the types and numbers of observations required in the associated compliance tests.

The eight papers in this volume of proceedings deal with diverse aspects of the statistical design and interpretation of environmental regulations and associated compliance monitoring programs. A unifying theme among them is that the statistical objectives and characteristics of the regulations should be specified right at the planning stage and should be drivers of the specific regulation designs rather than being (in)consequential afterthoughts.

Paul I. Feder
Chair, ASA/EPA Conference on Compliance Sampling
Battelle Memorial Institute

TABLE OF CONTENTS

Preface. JOHN C. BAILAR III, McGill University ii

Introduction. PAUL I. FEDER, Battelle Memorial Institute iii

Index of Authors vi

I. TOXICOKINETIC AND PERSONAL EXPOSURE CONSIDERATIONS IN THE DESIGN AND EVALUATION OF MONITORING PROGRAMS

Time Scales: Biological, Environmental, Regulatory. ALLAN H. MARCUS, Battelle Columbus Division 1

Discussion. RICHARD C. HERTZBERG, U.S. Environmental Protection Agency, ECAO–Cincinnati 16

Statistical Issues in Human Exposure Monitoring. WILLIAM C. NELSON, U.S. Environmental Protection Agency, EMSL–Research Triangle Park 17

Discussion. WILLIAM F. HUNT, JR., U. S. Environmental Protection Agency, OAQPS–Research Triangle Park 39

II. STATISTICAL DECISION AND QUALITY CONTROL CONCEPTS IN DESIGNING ENVIRONMENTAL STANDARDS AND COMPLIANCE MONITORING PROGRAMS

Designing Environmental Regulations. SOREN BISGAARD, WILLIAM G. HUNTER, University of Wisconsin–Madison 41

Discussion. W. BARNES JOHNSON, U.S. Environmental Protection Agency, OPPE–Washington, D.C. 51

Quality Control Issues in Testing Compliance with a Regulatory Standard: Controlling Statistical Decision Error Rates. BERTRAM PRICE, Price Associates, Inc. 54

Discussion. GEORGE T. FLATMAN, U.S. Environmental Protection Agency, EMSL–Las Vegas 75

III. COMPLIANCE WITH RADIATION STANDARDS

On the Design of a Sampling Plan to Verify Compliance with EPA Standards for Radium-226 in Soil at Uranium Mill Tailings Remedial-Action Sites. RICHARD O. GILBERT, Battelle Pacific Northwest Laboratory; MARK L. MILLER, Roy F. Weston, Inc.; H. R. MEYER, Chem–Nuclear Systems, Inc. 77

Discussion. JEAN CHESSON, Price Associates, Inc. 111

IV. THE BUBBLE CONCEPT APPROACH TO COMPLIANCE

Distributed Compliance: EPA and the Lead Bubble. JOHN W. HOLLEY, BARRY D. NUSSBAUM, U.S. Environmental Protection Agency, OMS–Washington, D.C. 112

Discussion. N. PHILIP ROSS, U.S. Environmental Protection Agency, OPPE–Washington, D.C. 121

V. COMPLIANCE WITH AIR QUALITY STANDARDS

Variable Sampling Schedules to Determine PM_{10} Status. NEIL H. FRANK, THOMAS C. CURRAN, U. S. Environmental Protection Agency, OAQPS–Research Triangle Park 122

Discussion. JOHN WARREN, U. S. Environmental Protection Agency, OPPE–Washington, D.C. 128

Analysis of the Relationship Between Maximum and Average in S02 Time Series. THOMAS S. HAMMERSTROM, Roth Associates, RONALD E. WYZGA, Electric Power Research Institute 129

Discussion. R. CLIFTON BAILEY, Health Care Financing Administration 154

Summary of Conference. JOHN C. BAILAR III, McGill University and U.S. Public Health Service 155

Appendix A: Program 160

Appendix B: Conference Participants 162

INDEX OF AUTHORS

Bailar, John C. ii,155
Bailey, R. Clifton 154
Bisgaard, Soren 41
Chesson, Jean 111
Curran, Thomas C. 122
Feder, Paul I. iii
Flatman, George T. 75
Frank. Neil H. 122
Gilbert, Richard O. 77
Hammerstrom, Thomas S. 129
Hertzberg, Richard C. 16
Holley, John W. 112
Hunt, Jr., William F. 39
Hunter, William G. 41
Johnson, W. Barnes 51
Marcus, Allan H. 1
Meyer, H. R. 77
Miller, Mark L. 77
Nelson, William C. 17
Nussbaum, B. D. 112
Price, Bertram 54
Ross, N. Philip 121
Warren, John 128
Wyzga, Ronald E. 129

TIME SCALES: BIOLOGICAL, ENVIRONMENTAL, REGULATORY

Allan H. Marcus
Battelle Columbus Division
P.O. Box 13758
Research Triangle Park, NC 27709

1. INTRODUCTION

E.P.A. has established primary air quality standards to protect the general public against the adverse health effects of air pollutants, and secondary standards to protect against other adverse environmental impacts. Compliance with these standards is usually prescribed by an explicit sampling protocol for the pollutant, with specified properties of the instrumentation and its calibration, appropriate location of the sampling device, and the frequency and averaging time of the samples. The temporal properties of the compliance sampling protocol represent a compromise among time scales of biological response to an environmental insult, variation in concentration to which the population is exposed, cost and precision of the sample data. Biological and health effects issues are primary and should be kept always in mind. Inadequate sampling schedules for compliance testing might allow fluctuating exposures of toxicological significance to escape detection. Resources for testing compliance are usually going to be scarce, and focusing on health effects may allow the analyst and designer of environmental regulations to find some path between oversampling and undersampling environmental data.

In this review I will emphasize air quality standards for lead. Lead is a soft dense metal whose toxic effects have long been known. In modern times atmospheric lead has become a community problem because of the large quantities of lead used as gasoline additives. While this problem was substantially reduced as a result of E.P.A.'s leaded gasoline phasedown regulations, there are still significant quantities of atmospheric lead around primary and secondary metal smelters, battery plants etc., and substantial residues of previous lead emissions in surface soil and dust. Other regulatory authorities control lead concentrations in drinking water, in consumer products, and in the work place. E.P.A.'s air lead regulations are spelled out in C.F.R. 40: 58 (1982). I will describe these in more detail below, along with some alternative approaches that are being considered.

I will also very briefly describe some of the biological and physical time scale problems arising in the effects of ozone on loss of agricultural crop yields. This will allow us to look at a gaseous pollutant whose effects include economic welfare as well as human health.

2. AIR LEAD STANDARDS

Atmospheric lead is largely found as inorganic lead salts on small particles, thus many of the data collection issues are similar to those encountered in sampling Total Suspended Particulates (TSP). A great deal of data has been collected by the State and Local Air Monitoring Stations (SLAMS) network. These provide information about areas where the lead concentration and population density are highest and monitoring for testing compliance with standards is most critical. In order for a SLAMS station to be part of the National Air Monitoring Station (NAMS) network, very specific criteria must be satisfied about sampler location in terms of height above ground level, distance from the nearest major roadway, and spatial scale of which the station is supposed to be representative. The citing study must also have a sufficiently long sampling period to exhibit typical wind speeds and directions, or a sufficiently large number of short periods to provide an average value consistent with a 24-hour exposure (CD, 1986).

The current averaging time for the lead primary National Ambient Air Quality Standard (NAAQS) is a calendar quarter (3 months), and the air lead NAAQS is a quarterly average of 1.5 ug/m3 that shall not be exceeded. The lead standard proposed in 1977 was based on an averaging time of one calendar month. The longer period has the advantage of greater statistical stability. However, the shorter period allows some extra protection. Clinical studies with adult male volunteer subjects showed that blood lead concentration (PbB) changed to a new equilibrium level after 2 or 3 months of exposure (Rabinowitz et al., 1973, 1976; Griffin et al., 1975). The shorter averaging time was also thought to give more protection to young children (42 FR 63076) even though there was no direct evidence then (or now!) on blood lead kinetics in children. "The risk of shorter term exposures to air lead concentrations elevated above a quarterly-averaged standard that might go undetected were considered in the 1978 standard decision to be minimized because 1) based on the ambient air quality data available at that time, the possibilities for significant, sustained excursions were considered small, and 2) it was determined that direct inhalation of air lead is a relatively small component of total airborne lead exposure (43 FR 46246)." (Cohen, 1986). The biological reasons for reevaluating the averaging time are discussed in the next section.

Alternative forms of the air lead standard are now being evaluated by E.P.A.'s Office of Air Quality Planning and Standards (OAQPS). The averaging time is only one of the components in setting an air lead standard. The "characterizing value" for testing compliance can assume a wide variety of forms, e.g. the maximum monthly (or quarterly) average as used in the "deterministic" form of the standards, the maximum of the average monthly mean over a specified number of years e.g. 3 consecutive years, the average of the maximum monthly averages for each year within a specified number of years, the average of the three highest months (or quarters) within a specified number of years etc. Some averaging of the extreme values certainly smoothes out the data, but also conceals extreme high-level excursions. Some attention has been given to the statistical properties of the alternative characterizing values (Hunt, 1986). The consequences of different characterizing values for biological exposure indices or health effects indicators has not yet been evaluated.

A final consideration is the sampling frequency. The current normal situation is a 24-hour average collected every 6th day. The number of samples collected also depends on the fraction of lost days: it is not uncommon for 25% of the data to be lost. Thus one might have only 3 or 4 valid samples per month. Hunt (1986) examined more frequent sampling schemes: every day, every other day, every third day. He also compared the consequences of deterministic vs. "statistical" form of the standard, monthly vs. quarterly characteristic values, 25% data loss vs. no loss. The community air lead problem in the U.S. is now more likely to be related to point sources than to area-wide emissions. thus the following three scenarios for location were evaluated: (1) source oriented sites with maximum annual quarterly averages less than 1.5 ug/m3; (2) source oriented sites with maximum annual quarterly average greater than 1.5 ug/m3; (3) NAMS urban maximum concentration sites. Some conclusions suggested by his study for quarterly averaging time are:

(i) The characterizing value with the best precision for specified sampling frequency is the statistical quarterly average.

(ii) The required sampling frequency could vary with site type. Respectively, (1) every other day for source oriented sites < 1.5 ug/m3; (2) every day for source oriented sites > 1.5 ug/m3; (3) for NAMS sites, every third day. The required precision here is +/-10% of the mean.

Hunt also found that more frequent sampling would be required if the monthly averaging times were used. The source-oriented sites would require every day sampling and the NAMS sites every-other-day sampling to achieve +/- 10% precision.

Is such intensive sampling actually required? Are we really interested in specified precision for atmospheric concentrations, or should we shift the focus of compliance sampling to more relevant indicators of biological effect? Let us examine some of these indicators.

3. BIOLOGICAL KINETICS OF LEAD

Lead is absorbed from the environment through the lungs (direct inhalation) and through the gastro-intestinal tract (ingestion). Organic compounds of lead may also be absorbed through the skin. Once lead is absorbed into blood plasma through the alveoli or through the gut lumen, it is quickly ionized and may henceforth regarded as indistinguishable by source. Thus the internal kinetics of lead may be deduced from experimental data whether lead uptake is by intravenous injection, inhalation or ingestion. Lead is distributed from plasma to the red blood cells, kidney, liver, skeleton, brain, and other tissues. The fractional absorption of lead from the plasma varies greatly from tissue to tissue, thus the time scales for transfer of lead also vary greatly. It is often assumed that lead equilibrates quickly and completely between plasma and red blood cells, thus the whole blood lead concentration can be used as a surrogate indicator of internal exposure. This is not the case.

The initial uptake of lead from plasma to the red blood cells is very rapid, occurring within a few minutes to tens of minutes (Campbell et al., 1984; Chamberlain, 1984; De Silva, 1981). Complete equilibration does not occur at all concentrations, however, since the relationship between whole blood lead and plasma lead becomes strikingly nonlinear at higher concentrations (Manton and Cook, 1985; Marcus, 1985a). The most plausible explanation is that there is reduced transfer of lead to the red blood cells at higher concentrations, whether attributed to reduced lead-binding capacity of the erythrocytes or reduced transfer rate across the erythrocyte membrane as lead concentrations increase. This is reinforced by multi-dose experiments on rats in which lead concentrations in brain, kidney, and femur are proportional to dose, which is expected if tissue concentrations equilibrate with plasma concentrations, not with whole blood lead concentrations.

Lead concentrations in peripheral tissues can be modeled by coupled systems of ordinary differential equations. Parameters for such systems can be estimated by iterative nonlinear least squares methods, often with Marquardt-type modifications to enlarge the domain of initial parameter estimates which allow convergence to the optimal solution (Berman and Weiss, 1978). Data sets with observations of two or more components often sustain indirect inferences about unobserved tissue pools. Analyses of data in (Rabinowitz et al., 1973, 1976; Griffin et al., 1975; De Silva, 1981) reported in (Marcus, 1985abc; Chamberlain, 1985; CD, 1986) show that lead is absorbed into peripheral tissues in adult humans within a few days. The retention of lead by tissues is much longer than is the initial uptake. Even soft tissues such as kidney and liver appear to retain lead for a month or so, and the skeleton retains lead for years or tens of years (Christoffersson et al., 1986).

The relevance of blood lead and tissue lead concentrations to overt toxicity is not unambiguous. As in any biologically variable population, some individuals can exhibit extremely high blood lead with only mild lead poisoning (Chamberlain and Massey, 1972). A more direct precursor of toxicity is the erythrocyte protoporphyrin (EP) concentration. Elevated levels of EP show that lead has deranged the heme biosynthetic pathway, reducing the rate of production of heme for hemoglobin. EP is now widely used as a screening indicator for potential toxicity. An example of the utility of EP is that after a brief massive exposure of a British worker (Williams, 1984), zinc EP increased to very elevated levels within a week of exposure even the worker was still largely asymptomatic. Even though there is considerable biological variability, EP levels in adults increase significantly within 10 to 30 days after beginning an experimental increase of ingested lead (Stuik, 1974; Cools et al., 1976; Schlegel and Kufner, 1978). Thus biological effects in adult humans occur very shortly after exposure, certainly within a month.

While the uptake of lead and the onset of potential toxicity occur rapidly during increased exposure, the reduction of exposure does not cause an equally rapid reduction in either body burden or toxicity indices. Accumulation of mobilizable pools of lead in the skeleton and other tissues create an endogenous source of lead that is only slowly eliminated. Thus the rapid uptake of lead during periods of increased exposure should be emphasized in setting standards for lead.

The experimental data cited above are indeed human data, but all for adults (almost all for males). We are not aware of any direct studies on lead kinetics in children. One of the more useful sets of data involves the uptake of lead by infants from formula and milk (Ryu et al., 1984, 1985). Blood lead levels and lead content of food were measured at 28 day intervals. The results are negative but informative: Blood lead levels in these infants appeared to equilibrate so much faster that no estimate of the kinetic parameters was possible. A very rough estimate by Duggan (1984) based on earlier input-output studies in infants (Ziegler et al., 1978) gave a blood lead half life (= mean life * log(2)) of 4 to 6 days. Duggan's method has many assumptions and uncertainties. An alternative method, allometric scaling based on surface area, suggests that if a 70 kg adult male has a blood lead mean life of 30 days, then a 7 kg infant should have a blood lead mean life of about 8 days.

The above estimates of lead kinetics in children are not strictly acceptable. Children are kinetically somewhat different from adults, with a somewhat larger volume of blood and much smaller but rapidly developing skeleton (especially dense cortical bone that retains most of the adult body burden of lead). Children also absorb lead from the environment at a greater rate, as they have greater gastrointestinal absorption of ingested lead and a more rapid ventilation rate than do adults. A biomathematical model has been developed by Harley and Kneip (1984) and modified for use by OAQPS. This uptake/biokinetic model is based on lead concentrations in infant and juvenile baboons, who are believed to constitute a valid animal model for human growth and development. Preliminary applications of the model are described by (Cohen, 1986; ATSDR, 1987; Marcus et al., 1987). The model includes annual changes of kinetic parameters such as the transfer rates for blood-to-bone, blood-to-liver, liver-to-gastrointestinal tract, and growth of blood, tissue, and skeleton. The model predicts a mean residence time for lead in blood of 2-year-old children as 8 days.

Blood lead concentrations change substantially during childhood (Rabinowitz et al., 1984). These changes reflect the washout of in utero lead, the exposure of the child to changing patterns of food and water consumption, and the exposure of the toddler to leaded soil and dust in his or her environment. We must thus consider also the temporal variations of exposure to environmental lead.

4. TIME SCALES OF LEAD EXPOSURE

Air lead concentrations change very rapidly, depending on wind speed and direction and on emissions patterns. Biological kinetics tend to filter out the "high-frequency" variations in environmental lead, so that only environmental variations on the order of a few days are likely to play much of a role. The temporal patterns depend on averaging time and sampling frequency, and thus will vary from one location to another depending on the major lead sources at that site. Figure 1 shows the time series for the logarithm of air lead concentration (log PbA) near a primary lead smelter in the northwestern U.S. The data are 24-hour concentrations sampled every third day (with a few minor slippages). We analysed these data using Box-Jenkins time series programs. The temporal structure is fairly complex, with a significant autoregressive component at lag 9 (27 days) and significant moving average components at lags 1 and 3 (3 days and 9 days). Time series analyses around point source sites and general urban sites may thus be informative.

Direct inhalation of atmospheric lead may be only a minor part of lead exposure attributable to air lead. Previously elevated air lead levels may have deposited a substantial reservoir of lead in surface soil and house dust in the environment; these are the primary pathways for air lead in children aged 1-5 years. Little is known about temporal variations in soil and house dust lead. Preliminary results cited in (Laxen et al., 1987) suggest that lead levels in surface dust and soil around redecorated houses and schools can change over periods of time of two to six months. While lead levels in undisturbed soils can persist for thousands of years, the turnover of lead in urban soils due to human activities is undoubtedly much faster.

Individuals are not stationary in their environment. Thus, the lead concentrations to which individuals are exposed must include both spatial and temporal patterns of exposure. The picture is complex, but much is being learned from personal exposure monitoring programs.

The amount of variation in air lead concentrations at a stationary monitor can be extremely large. Coefficients of variation in excess of 100% are not uncommon around point sources such as lead smelters, even when monthly or quarterly averages are used. This variability is far in excess of that attributable to meteorological variation and is due to fluctuations in the emissions process e.g. due to variations in feed stock, process control, or production rate. Furthermore, the concentration distributions are very skewed and heavy-tailed, more nearly log-normally distributed than normal even for long averaging times. The stochastic properties of the process are generally unknown, although it may be assumed that air, dust, and soil lead concentrations around point sources that have been in operation for a long time are approximately stationary. In most places in the United States, lead levels in all sources of exposure, including food, water, and paint, as well as those pathways from gasoline lead, have been declining. With these points in mind, we can begin to construct a quantitative characterization of a health effects target for compliance studies.

5. HEALTH EFFECTS CHARACTERIZATION: A THEORETICAL APPROACH

We will here briefly describe a possible approach to the problem of choosing an averaging time that is meaningful for health effects. Related problems such as sampling frequency then depend on the precision with which one wishes to estimate the health effects characterization. The basic fact is that all of the effects of interest are driven by the environmental concentration-exposure $C(t)$ at time t integrated over some period of time, with an appropriate weighing factor. As people are exposed to diverse pollutant sources, the uptake from all pathways must be added up. If the health effect is an instantaneous one whose value at time t is denoted $X(t)$, and if the biokinetic processes are all linear (as is assumed for OAQPS uptake/biokinetic model) or can be reasonably approximated by a linear model driven by $C(u)$ at time u, then the biokinetic model can be represented by an aftereffect term $f(t-u)$ after an interval $t-u$. Mathematically,

$$X(t) = \int f(t-u)\, C(u)\, du$$

The after effect function for linear compartmental models is a mixture of exponential terms.

The time-averaged concentration-exposure at time t, denoted $Y(t)$, is also a moving average of concentration $C(u)$ at time u, with a weight given by $g(t-u)$ after an interval t-u. Thus compliance will be based on the values of the variable $Y(t)$ in adjacent intervals, where

$$Y(t) = \int g(t-u)\, C(u)\, du$$

The simple time-weighted average for an averaging time of length T is

$$g(t-u) = 1/T \quad \text{if } t-T < u < t$$

$$= 0 \quad \text{otherwise.}$$

The properties of the moving average processes are easily evaluated, e.g. the expected value E[], variance [], covariance cov[.,.], are:

$$E[X(t)] = \int f(t-u)\, E[C(u)]\, du$$

$$var[X(t)] = \int\int f(t-u)\, f(t-v)\, cov[C(u),C(v)]\, du\, dv$$

$$cov[X(t),Y(s)] = \int\int f(t-u)\, g(s-v)\, cov[C(u),C(v)]\, du\, dv$$

Thus, we could formalize the problem of selecting an averaging time T by the following mathematical problem: choosing the averaging time T that maximizes the correlation between $X(t)$ and $Y(s)$, for that time t at which $E[X(t)]$ is maximum. That is, look for the time(s) t at which we expect the largest adverse health effect or effect indicator (e.g. blood lead). Then find the averaging time T such the moving average at some other time s is as highly correlated as possible with $X(t)$. Note that we do not require that $s = t$. We may also restrict the range of values of T.

EXAMPLE: ONE-COMPARTMENT BIOKINETIC MODEL, MARKOV EXPOSURE MODEL.

Suppose that the relevant biokinetic model is a simple one-compartment model. The aftereffect of a unit pollutant uptake is an exponential washout (e.g. of blood lead, to a first approximation) with time constant k,

$$f(t-u) = \exp(-k\,(t-u)) \quad \text{if } u < t$$

$$= 0 \quad \text{if } u > t$$

We will also assume that the concentration-exposure process $C(t)$ is stochastically second-order stationary with covariance function

$$cov[C(u),C(v)] = var[C]\, \exp(-a\, |u - v|)$$

After some algebra, one finds that:

$$\mathrm{var}[X(t)] = \mathrm{var}[C] / k(a + k)$$

$$\mathrm{var}[Y(t)] = \mathrm{var}[C]\, 2\, (aT-1 + \exp(-aT)) / a^2 T^2$$

If s-T < t < s then

$$\mathrm{cov}[X(t),Y(s)] = \mathrm{var}[C]\, [\, 2/ak + 2a \exp(-k(t+T-s))/k(k-a)(a+k) - \exp(-a(t+T-s))/a(k-a) - \exp(-a(s-t))/a(a+k)]/T$$

If t < s-T then

$$\mathrm{cov}[X(t),Y(s)] = \mathrm{var}[C]\, [\exp(-a(s-t-T))-\exp(-a(s-t))]/Ta(a+k)$$

If t > s (for predicting from the current sampling time s to future time t) then

$$\mathrm{cov}[X(t),Y(s)] = \mathrm{var}[C]\, [\exp(-a(t-s))(1-\exp(-aT))/a(k-a) - 2a \exp(-k(t-s))(1-\exp(-kT))/k(a+k)(k-a)]/T$$

A small table of correlations between X(t) and Y(t) are shown in Table 1 for an assumed averaging time T=30 days. Note that the correlations between fluctuations in blood lead concentration X(t) and monthly averaged lead concentration Y(t) are fairly high, but much worse for children than for adults when environmental concentrations fluctuate rapidly. These correlations are long-term averages for one subject: the correlation in real populations will be greatly attenuated due to differences in biological parameters and exposures among people.

TABLE 1

Correlation between Blood Lead and Monthly Average Lead

Environmental Lead Time Constant a	Blood Lead Kinetic Parameter k	
	1/(8 d) Child	1/(40 d) Adult
1/(4 d)	0.7707	0.8783
1/(10 d)	0.8476	0.8933
1/(25 d)	0.9236	0.9132

The uses of this method for assessing the relationship between health effects and averaging time are shown in Table 2 for the sensitive case of rapid fluctuations in air lead concentration. It is clear that for this simple model, the averaging time T with highest correlation for children or for adults is about 1.5/k, and that much longer or much shorter averaging times will not capture significant excursions in blood lead. An averaging time of 15-20 days will make Y(t) reasonably predictive of X(t) for both adults and children.

TABLE 2

CORRELATION BETWEEN BLOOD LEAD CONCENTRATION AND AVERAGE ENVIRONMENTAL LEAD CONCENTRATION AS A FUNCTION OF AVERAGING TIME T

Assumed environmental lead correlation scale a = 1/(4 days)

	CORRELATION	
Averaging Time T, Days	CHILD k = 1/(8 days)	ADULT k = 1/(40) days)
5	0.8792	0.5062
7	0.9287	0.5674
10	0.9588	0.6430
14	0.9497	0.7207
20	0.8900	0.8020
30	0.7707	0.8783
60	0.5451	0.9141
90	0.4402	0.8579

Samples collected for compliance testing have a more complicated structure for the weight function g(t-u), namely (for h-hour samples once every m days in an internal of T days).

$g(t-u) = m/hT$ if $t_o + (j-1)H < u < t_o + (j-1)(H+h)$

where H = 24m hours

t_o = beginning of last compliance interval before t

$j = 1, \ldots, m$

and $g(t-u) = 0$ otherwise

That is, g(t-u) is the sum of T/m rectangles spaced H-hours apart. Similar calculations could be done using this g(t).

Assessment of realistic situations will require careful attention to both the biokinetic model represented by f(t), and the temporal variations in exposure represented by cov(C(u),C(v)] etc. The example represented above is the simplest representation of the interplay of biological time scales (represented by k), environmental time scales (represented by a), and regulatory time scales (represented by T). Numerical evaluation of realistic examples should proceed as above. If the underlying biokinetic model is severely nonlinear, then computer simulations will be needed. The concentration-exposure function here subsumes all spatial variation. Realistic human exposure models to various microenvironments may be needed as well. Thus the function C(t) here is a composite, including fractional absorption of environmental lead, volume of environmental intake (e.g. m^3/d of air, L/d of water, mg/d of leaded soil and dust, g/d of food) as well as concentration C(t).

6. TIME SCALES FOR THE EFFECTS OF OZONE ON AGRICULTURAL CROP YIELDS

The regulation of ozone has for some time been one of E.P.A.'s most pressing problems -- a regulatory irritant as well as a lung irritant. The secondary standards for ozone have drawn considerable attention, due to the knowledge that exposure to ozone may cause economically significant damage to cash crops and forests. The time of day of the ozone exposure, and the day of exposure during the growing season, may seriously determine the effects of exposure and consequently of the statistics that are used to formulate the standard. A number of approaches to defining a biologically relevant standard are being investigated (Lee et al, 1987ab; Larsen et al., 1987).

Air monitoring data have been collected in connection with the chamber studies of the National Crop Loss and Assessment Network (NCLAN) and related studies have been carried out at E.P.A.'s Corvallis Environmental Research Laboratory (CERL). The earlier NCLAN data were based on seven hours of monitoring (0900-1600) and statistics appropriate to that period. More recent studies use longer sampling periods, including 24-hour samples at CERL. Examples of the time patterns of exposure used at CERL are shown in Lee et al., 1997ab. The characterizations of the air monitoring data considered for use as exposure statistics and compliance specifications include the following, all based on the mean hourly ozone concentration C(h) at hour h:

MEAN STATISTICS

M7 = seasonal mean of C(h) for 0900-1600 hr each day

M1 = seasonal mean of daily maximum C(h) during 7 hours

Effective Mean = ($ C(h)**p /N) **1/p [Note: $ means sum]

PEAK STATISTICS

P7 = seasonal peak of 7-hour daily mean over 0900-1600 hrs.

P1 = seasonal peak hourly concentration

CUMULATIVE STATISTICS

Total Exposure = $ C(h)

Total Impact = ($ C(h)**p)**1/p

Phenologically Weighted Cumulative Impact (PWCI)

= ($ C(h)**p w(h))**1/p

EXCEEDANCE STATISTICS

HRSxx = number of hours in which C(h) > .xx ppm ozone

SUMxx = total ozone concentration X hours with C(h) .xx

and at least six other statistics characterizing episode lengths etc.
The statistic most frequently considered for ozone characterization is M7. However, the statistics that best predict crop loss as a function of exposure are peak-weighted cumulative measures (not averaged by the number of hours in the growing season) such as Total Impact, SUMxx, or PWCI. The PWCI index allows different weight to be placed at different times in the growing season. For example, the dry shoot weight of three cuttings of alfalfa in a CERL experiment was transformed to a fraction of the controls. The values of M7 clearly measure the damaging effect of ozone, but with a great deal of scatter around the regression line. The somewhat clustered values of M7 are spread out by the statistic PWCI that gives much higher weight to large values of C(h) (as C(h)**2) and to the most recent ozone exposures (weight 1 to the most recent exposures, weight 0.8 to those preceding the previous cutting, and weight 0.1 to those preceding the next earlier cutting). Crop loss is much better defined by the values of PWCI, with relatively little scatter about the fitted curve of "Weibull" form.

The ozone example suggests that biological time scales of response are better captured by compliance statistics that give higher weight to recent exposures, as in our lead example. However, the biokinetics are clearly nonlinear in ozone concentration so that some noncompartmental mechanism of damage, repair, and metabolism must be assumed to be operating. The PWCI is a cumulative value and not a peak or exceedance statistic, thus even low levels of ozone exposure appear to be causing some damage. The biological statistic for compliance sampling (for alfalfa, anyway) is thus a 24-hour peak-weighted cumulative exposure statistic. The one-hour averaging time appears appropriate.

7. CONCLUSIONS

The variables that are used to formulate pollution standards and determine compliance with those standards are usually defined in terms of moving averages of "instantaneous" concentrations. In this paper we have shown that weighted moving averages of concentrations, closer to predictive models of biological effects or indicators of effects, are sometimes also moving averages of concentrations. Thus certain aspects of monitoring and compliance sampling (e.g. averaging time and sampling frequency) could be evaluated in terms of the correlation of the compliance statistic with predicted biological effects, and the precision with which the predicted biological effect (not the compliance statistics) could be measured. Thus there are some pollutants for which compliance sampling could be tied more directly to models of health effects and biological damage, providing E.P.A. with an inexpensive methodology for assessing potential risks to exposed populations. These methods may also be used to assess the likelihood that loose compliance sampling schedules will allow excursions of high pollutant concentration that are potentially toxic.

For most chemicals of interest there is not nearly enough information on pharmacokinetics, toxicokinetics, or temporal variability of exposure pattern to allow these calculations to be made. However, for many criteria pollutants, the level of information is adequate and the ratio between typical population levels so close to a health effects criterion level as to make this a serious issue. For example, in 1978, the criterion level for blood lead was 30 ug/dl, but the geometric mean blood lead in urban children was about 15 ug/dl, of which 12 ug/dl was assumed to be "non-air" background (i.e. regulated by some other office). Due to the reduction of leaded gasoline during the 1970's, the mean blood lead level for urban children had fallen to 9-10 ug/dl by 1980, and is likely to be somewhat lower today. However, better data on health effects (e.g. erythrocyte protoporphyrin increases in iron-deficient children or hearing loss and neurobehavioral problems) in children with lead burdens now suggest a much lower health criterion level is appropriate, perhaps 10-15 ug/dl. Thus there is still very little "margin of safety" against random excursions of lead exposure.

This is also true for other criteria pollutants, especially for sensitive or vulnerable subpopulations. For example, asthmatics may experience sensitivity to elevated levels of sulfur dioxide or ozone, especially when exercising. Activity levels certainly affect the kinetics of gaseous pollutant uptake and elimination. Subpopulation variations in kinetics and pharmacodynamics may be important. Acute exposure sampling in air or water (e.g. 1-day Health Advisories for drinking water) should be sensitive to pharmacokinetic time scales.

Biokinetic information on pollutant uptake and metabolism in humans is not often available for volatile organic compounds and for most carcinogens. Thus large uncertainty factors for animal extrapolation and for route of exposure variations are used to provide a conservative level of exposure. The methods shown here maybe less useful in such situations. But the development of realistic biologically motivated pharmacokinetic models for extrapolating animal data to humans may establish a larger role for assessment of compliance testing for these substances.

ACKNOWLEDGEMENTS

I am grateful to Ms. Judy Kapadia for retyping the manuscript, and to the reviewer for his helpful comments.

REFERENCES

Berman M, Weiss MF. 1978. SAAM - Simulation, Analysis, and Modeling. Manual. U.S. Public Health Service Publ. NIH-180.

Campbell BC, Meredith PA, Moore MR, Watson WS. 1984. Kinetics of lead following intravenous administration in man. Tox Letters 21:231-235.

CD [Criteria Document]. 1986. Air quality criteria for lead. Environmental Criteria and Assessment Office, US Environmental Protection Agency. EPA-600/8-83/028aF (4 volumes). Res. Tri. Pk., NC.

Chamberlain AC. 1985. Prediction of response of blood lead to airborne and dietary lead from volunteer experiments with lead isotopes. Proc Roy Soc Lond B224:149-182.

Chamberlain MJ, Massey PMO. 1972. Mild lead poisoning with excessively high blood lead. Brit J Industr Med 29:458-461.

Christoffersson JO, Ahlgren L, Schutz A, Skerfving S. 1986. Decrease of skeletal lead levels in man after end of occupational exposure. Arch Env Health 41:312-318.

Cohen, J. Personal communications about OAQPS staff paper, April-Nov. 1986.

Cools A, Salle JA, Verberk MM, Zielhuis RL. 1976. Biochemical response of male volunteers ingesting inorganic lead for 49 days. Int Arch Occup Environ Health 38:129-139.

DeSilva PE. 1981. Determination of lead in plasma and studies on its relationship to lead in erythrocytes. Brit J Industr Med 38:209-217.

Duggan MJ. 1983. The uptake and excretion of lead by young children. Arch Environ Health 38:246-247.

Griffin TB, Coulston F, Wills H, Russell JC, Knelson JH. 1975. Clinical studies of men continuously exposed to airborne lead. Environ Quality Safety Suppl 2:254-288.

Harley NH, Kneip TH. 1985. An integrated metabolic model for lead in humans of all ages. Report, New York Univ. Dept. Environ. Med.

Hunt, WF Jr. 1986. A comparison of the precision associated with alternative sampling plans for one versus three years of information and monthly versus quarterly averaging times. Memorandum to John Haines. Office of Air Quality Planning and Standards, US Environ. Protect. Agency. Jan. 30, 1986.

Larsen RI, Heck WW. 1984. An air quality data analysis system for interrelating effects, standards, and needed source reductions: Part 8. An effective mean O3 crop reduction mathematical model. J Air Pollut Control Assoc 34: 1023-1034.

Larsen RI, McCurdy TR, Johnson PM. 1987. The relative importance of ozone concentration and its variation in injuring soybean. Draft report, Atm. Sci. Res. Lab., US Env. Protection Agency, Res. Tri. Pk. NC

Laxen, DPH, Lindsay F, Raab GM, Hunter R, Fell GS, Fulton M. 1987. The variability of lead in dusts within the homes of young children. In Lead In the Home Environment, ed. E. Culbard. Science Services. London.

Lee EH, Tingey DT, Hogsett WE. 1987a. Selection of the best exposure-response model using various 7-hour ozone exposure statistics. Report for Office of Air Quality Planning and Standards, US Environ. Protection Agency.

Lee EH, Tingey DT, Hogsett WE. 1987b. Evaluation of ozone exposure statistics in exposure-response relationships. Submitted for publication.

Manton WI, Cook JD. 1984. high accuracy (stable isotope dilution) measurements of lead in serum and cerebrospinal fluid. Brit J Industr Med 41:313-319.

Marcus AH. 1985a. Multicompartment kinetic models for lead. Part III. Lead in blood plasma and erythrocytes. Environ Res 36: 473-489.

Marcus, AH. 1985b. Multicompartment kinetic models for lead. Part II. Linear kinetics and variable absorption in humans without excessive lead exposures. Environ Res 36: 459-472.

Marcus, AH. 1985c. Testing alternative nonlinear kinetic models in compartmental analysis. In: Eisenfeld J, DeLisi C, eds. Mathematics and Computers in Biomedical Applications, Elsevier Science. New York, pp. 259-268.

Rabinowitz M, Leviton A, Needleman H. 1984. Variability of blood lead concentrations during infancy. Arch Environ Health 39:74-77.

Rabinowitz MB, Wetheril GW, Kopple JD. 1973. Lead metabolism in the normal human: stable isotope studies. Science 162:725-727.

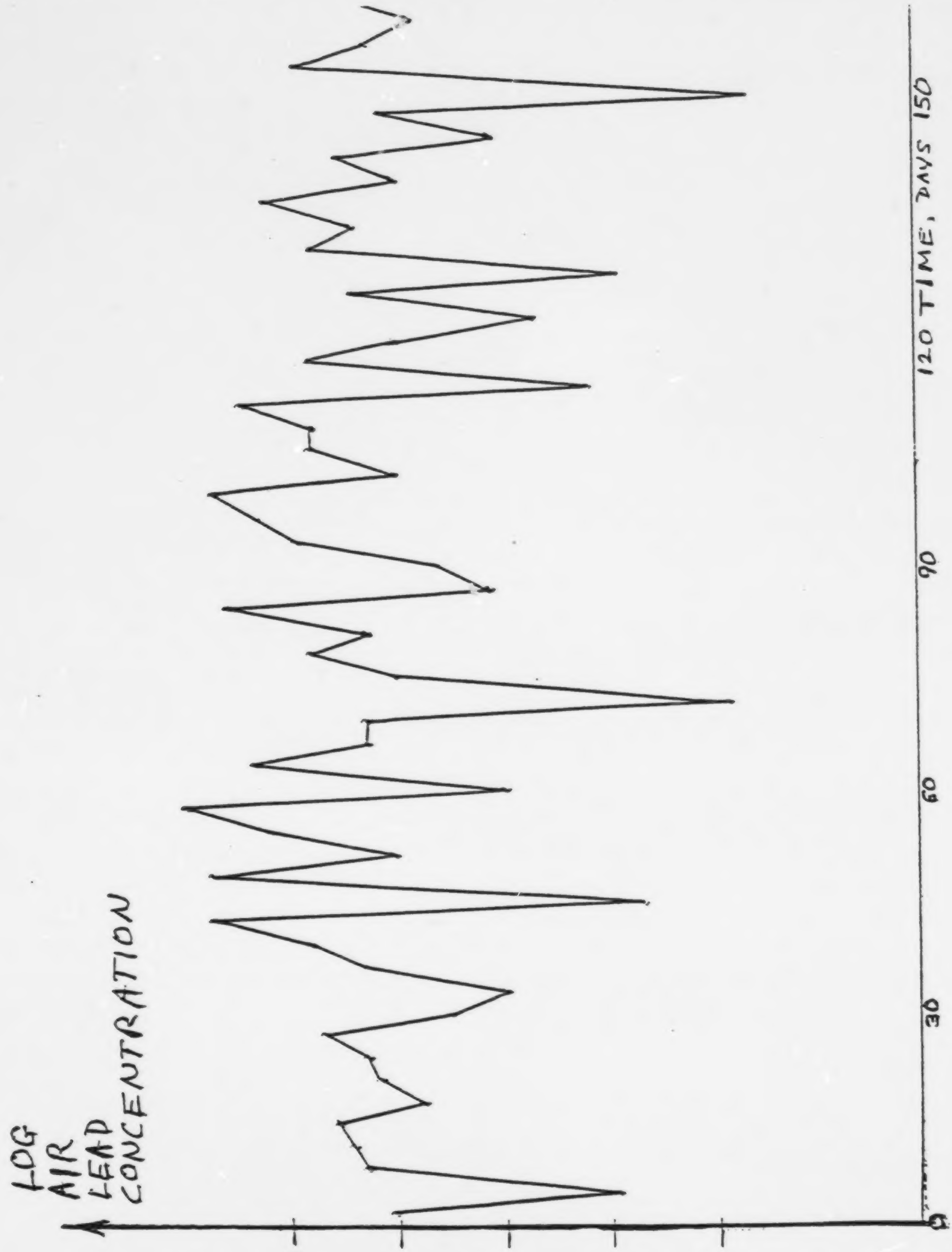
Rabinowitz MB, Wetherill GW, Kopple JD. 1976. Kinetic analysis of lead metabolism in healthy humans. J Clin Invest 58:260-270.

Ryu JE, Ziegler EE, Nelson SE, Fomon SJ.Dietary and environmental exposure to lead and blood lead during early infancy. In Dietary and Environmental Lead: Human Health Effects, ed Mahaffey K. Elsevier Science. New York, pp. 187-209.

Schlegel H, Kufner G. 1979. Long-term observation of biochemical effects of lead in human experiments. J Clin Chem Clin Biochem 17:225-233.

Stuik EJ. 1974. Biological response of male and female volunteers to inorganic lead. Int Arch Arbeitsmed 33:83-97.

Williams MK. 1984. Biological tests of lead absorption following a brief massive exposure. J Occup Med 26:532-533.


LOG AIR LEAD CONCENTRATION
4
2
1
.5
0.25
0
30
60
90
120 TIME, DAYS 150

DISCUSSION

Richard C. Hertzberg
Environmental Criteria and Assessment Office, U.S. EPA, Cincinnati, OH 45268

Comments on
"Time Scales: Biological, Environmental, Regulatory," Allan H. Marcus

Summary of Presentation

Marcus presents a case for consideration of physiologic time scales in the determination of compliance sampling protocols. The general theme of incorporating physiologic time into risk assessment is certainly scientifically supportable (e.g., NAS Workshop, 1986, "Pharmacokinetics in Risk Assessment," several authors), but has been previously proposed only for setting standards. Marcus takes the application one step further by showing how improper sampling can fail to detect exposure fluctuations that have toxicological significance.

The Regulatory Context

The modeling and data that Marcus presents seem reasonable, but key items seem to be missing, at least if this approach is to become used by regulatory agencies. The examples should show that the refinement will make a practical difference in the "cost-benefit" evaluation, and that the required data are accessible.

The first question is: does it matter? Most standards are set with a fair degree of conservatism, so that slight excursions above the standard will not pose a significant health risk. The first impression of Marcus' proposal is that it is fine tuning, when in fact it is the coarse control which needs to be turned. Let us consider the example of lead. Recent research has suggested that significant impairment of neurological development can be caused by lead concentrations much lower than previously thought. In fact, some scientists have suggested that lead toxicity may be a no-threshold phenomenon. If such is the case, then EPA's approach to setting lead standards will change drastically, and Marcus' example, though not necessarily his proposal, will probably not apply. But even with the current standard, it is not clear that results from Marcus' method will not be lost in the usual noise of biological data. For example, consider his figure showing the graphs of data and model fits for 11 human subjects. First, these results may be irrelevant to the air pollution issue since that data are following "ingestion" of lead, not "inhalation." Lead inhalation is in many ways more complicated than ingestion. Also, using day 30 as an example, the fitted erythrocyte protoporphyrin levels vary dramatically across individuals (mean=49, s.d.=20.3, range=30–73). I could not read the graphs well, but even accounting for differing starting values, the curve shapes also change across individuals, so that predictions for any untested individual might be difficult.

The second question, that of data requirements, cannot be answered from this presentation alone. But some issues can be mentioned. It is not clear that the correlations between blood lead (Table 1) and monthly average lead are good predictors of the correlation between monthly average lead and neurological impairment. But is the correlation the best indicator of performance? A better question, perhaps, is: do changes in blood lead which could be allowed by using the weakest sampling protocol actually result in significantly increased incidence of neurological dysfunction, when compared to the best compliance sampling procedure as determined using Marcus' scheme? It is not clear how much data would be required to answer that question.

Also, it seems that Marcus' approach must have pharmacokinetic data on humans. The data requirements are then more severe for most of the thousands of environmental chemicals, where only animal data are available. The situation is even worse for carcinogens, where human cancer incidence data are not available at the low regulatory levels. In fact, the orders-of-magnitude uncertainty in the low-dose extrapolation of cancer bioassays easily swamps the error due to non-optimal compliance sampling.

So where might this research go? Certainly it should be further developed. This approach will definitely be useful for acute regulatory levels, such as the 1-day Health Advisories for drinking water, where internal dose and toxicity are closely tied to pharmacokinetics. It will probably be more significant for sensitive subgroups, such as children and those with respiratory disease, where the pharmacokinetics are likely to be much different from the norm, and where the tolerance to chemical exposure is already low. For those cases, scaling factors and uncertainty factors are highly inaccurate. Most important is the example Marcus presents, chemicals where uptake and elimination rates are dramatically different. For control of those chemicals, using the "average" monitored level is clearly misleading, and some approach such as Marcus' must be used. I would recommend the following steps:

- First, demonstrate the need. List at least a few chemicals that are being improperly monitored because of their pharmacokinetic properties.
- Then, show us that your method works and is practical.

Statistical Issues in Human Exposure Monitoring

William C. Nelson, U.S. EPA, EMSL, Research Triangle Park

ABSTRACT

Pollutant exposure information provides a critical link in risk assessment and therefore in environmental decision making. Traditionally, outdoor air monitoring stations have been necessarily utilized to relate air pollutant exposures to groups of nearby residents. This approach is limited by (1) using only the outdoor air as an exposure surrogate when most individuals spend relatively small proportions of time outdoors and (2) estimating exposure of a group rather than an individual. More recently, air monitoring of non-ambient locations, termed microenvironments, such as residences, offices, and shops has increased. Such data when combined with time and activity questionnaire information can provide more accurate estimates of human exposure. Development of portable personal monitors that can be used by the individual study volunteer provides a more direct method for exposure estimation. Personal samplers are available for relatively few pollutants including carbon monoxide and volatile organic compounds (VOC's) such as benzene, styrene, tetrachloroethylene, xylene, and dichlorobenzene. EPA has recently performed carbon monoxide exposure studies in Denver, Colorado and Washington, D.C. which have provided new information on CO exposure for individual activities and various microenvironments. VOC personal exposure studies in New Jersey and California have indicated that, for some hazardous chemicals, individuals may receive higher exposure from indoor air than from outdoor air. Indoor sources include tobacco smoke, cleansers, insecticides, furnishings, deodorizers, and paints. Types of exposure assessment included in these studies are questionnaires, outdoor, indoor, personal, and biological (breath) monitoring.

As more sophisticated exposure data become available, statistical design and analysis questions also increase. These issues include survey sampling, questionnaire development, errors-in-variables situation, and estimating the relationship between the microenvironment and direct personal exposure. Methodological development is needed for models which permit supplementing the direct personal monitoring approach with an activity diary which provides an opportunity for combining these data with microenvironment data to estimate a population exposure distribution. Another situation is the appropriate choice between monitoring instruments of varying precision and cost. If inter-individual exposure variability is high, use of a less precise instrument of lower cost which provides an opportunity for additional study subjects may be justified. Appropriate choice of an exposure metric also requires more examination. In some instances, total exposure may not be as useful as exposure above a threshold level.

Because community studies using personal exposure and microenvironmental measurements are expensive, future studies will probably use smaller sample sizes but be more intensive. However, since such studies provide exposure data for individuals rather than only for groups, they may not necessarily have less statistical power.

INTRODUCTION

Pollutant exposure information is a necessary component of the risk assessment process. The traditional approach to investigating the relationship between pollutant level in the environment and the concentration available for human inhalation, absorption or ingestion, has been 1) measurements at an outdoor fixed monitoring site or 2) mathematical model estimates of pollutant concentration from effluent emission rate information.[1]

The limitations of such a preliminary exposure assessment have become increasingly apparent. For example, recognition of the importance of indoor pollutant sources, particularly considering the large amount of time spent indoors, and concern for estimating total personal exposure have lead to more in-depth exposure assessments.

One of the major problems to overcome when conducting a risk assessment is the need to estimate population exposure. Such estimates require information on the availability of a pollutant to a population group via one or more pathways. In many cases, the actual concentrations encountered are influenced by a number of parameters related to activity patterns. Some of the more important are: the time spent indoors and outdoors, commuting, occupations, recreation, food consumption, and water supply. For specific situations the analyses will involve one major pathway to man (e.g. outside atmospheric levels for ozone), but for others, such as heavy metals or pesticides, the exposure will be derived from several different media.

A framework for approaching exposure assessments for air pollutants has been described by the National Academy of Science Epidemiology of Air Pollution Committee.[2] The activities shown in Figure 1 were considered to be necessary to conduct an in-depth exposure assessment.

As knowledge about the components of this framework, particularly sources and effects, has increased, the need for improved data on exposures and doses has become more critical. A literature review published in 1982 discussed a large number of research reports and technical papers with schemes for calculating population exposures.[3] However, such schemes are imperfect, relying on the limited data available from fixed air monitoring stations and producing estimates of "potential exposures" with unknown accuracy. Up until the 1980's, there were few accurate field data on the actual exposures of the population to important environmental pollutants. Very little was known about the variation from person to person of exposure to a given pollutant, the reason for these variations, or the differences in the exposures of subpopulations of a city. Furthermore, a variety of field studies undertaken in the 1970s and early 1980s showed that the concentrations experienced by people engaged in various activities (driving, walking on sidewalks, shopping in stores, working in buildings, etc.) did not correlate well with the simultaneous readings observed at fixed air-monitoring stations.[4-9] Two reviews have summarized much of the literature on personal exposures to environmental pollution showing the difficulty of relating conventional outdoor monitoring data to actual exposures of the population.[10,11] No widely acceptable methodology was available for predicting and projecting future exposures of a population or for estimating how population exposures might change in response to various regulatory actions. No satisfactory exposure framework or models existed.

TOTAL HUMAN EXPOSURE

The total human exposure concept seeks to provide the missing component in the full risk model: estimates of the total exposures of the population to environmental pollutants, with known accuracy and precision. Generating this new type of information requires developing an appropriate research program and methodologies. The methodology has been partially developed for carbon monoxide (CO), volatile organic compounds (VOC's) and pesticides, and additional research is needed to solve many problems for a variety of other pollutants.

The total human exposure concept defines the human being as the target for exposure. Any pollutant in a transport medium that comes into contact with this person, either through air, water, food, or skin, is considered to be an exposure to that pollutant at that time.

The instantaneous exposure is expressed quantitatively as a concentration in a particular carrier medium at a particular instant of time, and the average exposure is the average of the concentration to the person over some appropriate averaging time. Some pollutants, such as CO, can reach humans through only one carrier medium, the air route of exposure. Others, such as lead and chloroform, can reach humans through two or more routes of exposure (e.g., air, food, and water). If multiple routes of exposure are involved, then the total human exposure approach seeks to determine a person's exposure (concentration in each carrier medium at a particular instant of time) through all major routes of exposure.

Once implemented, the total human exposure methodology seeks to provide information, with known precision and accuracy, on the exposures of the general public through all environmental media, regardless of whether the pathways of exposure are air, drinking water, food, or skin contact. It seeks to provide reliable, quantitative data on the number of people exposed and their levels of exposures, as well as the sources or other contributors responsible for these exposures. In the last few years, a number of studies have demonstrated these new techniques. The findings have already had an impact on the Agency's policies and priorties. As the methodology evolves, the research needs to be directed toward identifying and better understanding the nation's highest priority pollutant concerns.

The major goals of the Total Human Exposure Program can be summarized as follows:

- Estimate total human exposure for each pollutant of concern
- Determine major sources of this exposure
- Estimate health risks associated with these exposures
- Determine actions to eliminate or at least reduce these risks

The total human exposure concept considers major routes of exposure by which a pollutant may reach the human target. Then, it focuses on those particular routes which are relevant for the pollutants of concern, developing information on the concentrations present and the movement of the pollutants through the exposure routes. Activity information from diaries maintained by respondents helps identify the microenvironments of greatest concern, and in many cases, also helps identify likely contributing sources. Biological samples of body burden may be measured to confirm the exposure measurements and to estimate a later step in the risk assessment framework.

In the total human exposure methodology, two complementary conceptual approaches, the direct and the indirect, have been devised for providing the human exposure estimates needed to plan and set priorities for reducing risks.

Direct Approach

The "direct approach" consists of measurements of exposures of the general population to pollutants of concern.[12] A representative probability based sample of the population is selected based on statistical design. Then, for the class of pollutants under study, the pollutant concentrations reaching the persons sampled are measured for the relevant environmental media. A sufficient number of people are sampled using appropriate statistical sampling techniques to permit inferences to be drawn, with known precision, about the exposures of the larger population from which the sample has been selected. From statistical analyses of subject diaries which list activities and locations visited, it usually is possible to identify the likely sources, microenvironments, and human activities that contribute to exposures, including both traditional and nontraditional components.

To characterize a population's exposures, it is necessary to monitor a relatively large number of people and to select them in a manner that is statistically representative of the larger population. This approach combines the survey design techniques of the social scientist with the latest measurement technology of the chemist and engineer, using both statistical survey methodology and environmental monitoring in a single field survey. It uses the new miniaturized personal exposure monitors (PEMs) that have become available over the last decade,[13,14,15] and it adopts the survey sampling techniques that have been used previously to measure public opinion and human behavior. The U.S. EPA Office of Research and Development (ORD) has recently conducted several major field studies using the direct approach, namely, the Total Exposure Assessment Methodology (TEAM) Study of VOCs, the CO field studies in Washington, D.C. and Denver, and the non-occupational exposure to pesticides study. These studies will be described later.

Indirect Approach

Rather than measuring personal exposures directly as in the previous approach, the "indirect approach" attempts to construct the exposure profile mathematically by combining information on the times people spend in particular locations (homes, automobiles, offices, etc.) with the concentrations expected to occur there. This approach requires a mathematical model, information on human activity patterns, and statistical information on the concentrations likely to occur in selected locations, or "microenvironments".[16] A microenvironment can be defined as a location of relatively homogeneous pollutant concentration that a person occupies for some time period. Examples include a house, office, school, automobile, subway or bus. An activity pattern is a record of time spent in specific microenvironments.

In its simplest form the "indirect approach" seeks to compute the integrated exposure as the sum of the individual products of the concentrations encountered by a person in a microenvironment and the time the person spends there. The integrated exposure permits computing the average exposure for any averaging period by dividing the time duration of the averaging period. If the concentration within microenvironment j is assumed to be constant during the period that person i occupies microenvironment j, then the integrated exposure E_i for the person i will be the sum of the product of the concentration c_j in each microenvironment and the time spent by person i in that microenvironment

$$E_i = \sum_{j=1}^{J} c_j t_{ij},$$

where E_i = integrated exposure of person i over the time period of interest;

c_j = concentrations experienced in microenvironment j;

t_{ij} = time spent by person i in microenvironment j; and

J = total number of microenvironments occupied by person i over the time period of interest.

To compute the integrated exposure E_i for person i, it obviously is necessary to estimate both c_j and t_{ij}. If T is the averaging time, the average exposure $\overline{E}_i$ of person i is obtained by dividing by T; that is $\overline{E}_i = E_i/T$, where E_i is summed over time T.

Although the direct approach is invaluable in determining exposures and sources of exposure for the specific population sampled, the Agency needs to be able to extrapolate to much larger populations. The indirect approach attempts to measure and understand the basic relationships between causative variables and resulting exposures, usually in particular microenvironments, through "exposure modeling." An exposure model takes data collected in the field, and then, in a separate and distinct activity, predicts exposure. The exposure model is intended to complement results from direct studies and to extend and extrapolate these findings to other locales and other situations. Exposure models are not traditional dispersion models used to predict outdoor concentrations; they are different models designed to predict the exposure of a rather mobile human being. Thus, they require information on typical activities and time budgets of people, as well as information on likely concentrations in places where people spend time.

The U.S. EPA ORD has also conducted several studies using the indirect approach. An example of a recent exposure model is the Simulation of Human Activities ad Pollutant Exposures (SHAPE) model, which has been designed to make predictions of exposures to population to CO in urban areas. This model is similar to the NAAQS Exposure Model (NEM). The SHAPE model used the CO concentrations measured in the Washington-Denver CO study to determine the contributions to exposure from commuting, cooking, cigarette smoke, and other factors. Once a model such as SHAPE is successfully validated (by showing that it accurately predicts exposure distributions measured in a TEAM field study), it can be used in a new city without a field study to make a valid prediction of that population's exposures using that city's data on human activities, travel habits, and outdoor concentrations. The goal of future development is to apply the model to other pollutants (e.g., VOCs, household pesticides) making it possible to estimate exposure frequency distributions for the entire country, or for major regions.

Field Studies

The total human exposure field studies from a central part of the U.S. EPA ORD exposure research program. Several studies have demonstrated the feasibility of using statistical procedures to choose a small representative sample of the population from which it is possible to make inferences about the whole population. Certain subpopulations of importance from the standpoint of their unique exposure to the pollutant under study are "weighted" or sampled more heavily than others. In the subsequent data analysis phases, sampling weights are used to adjust for the overrepresentation of these groups. As a result, it is possible to draw conclusions about the exposures of the larger population of a region with a study that is within acceptable costs.

Once the sample of people has been selected, their exposures to the pollutant through various environmental media (air, water, food, skin) are measured. Some pollutants have negligible exposure routes through certain media, thus simplifying the study. Two large-scale total human exposure field studies have been undertaken by U.S. EPA to demonstrate this methodology: the TEAM study of VOCs and the Denver - Washington DC, field study of CO.

The first set of TEAM Studies (1980-84) were the most extensive investigation of personal exposures to multiple pollutants and corresponding body burdens. In all, more than 700 persons in 10 cities have had their personal exposures to 20 toxic compounds in air and drinking water measured, together with levels in exhaled breath as an indicator of blood concentration.[17-19] Because of the probability survey design used, inferences can be made about a larger target population in certain areas: 128,000 persons in Elizabeth/Bayonne, NJ; 100,000 persons in the South Bay Section of Los Angeles, CA; and 50,000 persons in Antioch/Pittsburg, CA.

The major findings of the TEAM Study may be summarized as follows:

1. Great variability (2-3 orders of magnitude) of exposures occur even in small geographical areas (such as a college campus) monitored on the same day.

2. Personal and overnight indoor exposures consistently outweigh outdoor concentrations. At the higher exposure levels, indoor concentrations may be 10-100 times the outdoor concentrations, even in New Jersey.

3. Drinking water and beverages in some cases are the main pathways of exposure to chloroform and bromodichloromethane -- air is the main route of exposure to 10 other prevalent toxic organic compounds.

4. Breath levels are significantly correlated with previous personal air exposures for all 10 compounds. On the other hand, breath levels are usually not significantly correlated with outdoor levels, even when the outdoor level is measured in the person's own backyard.

5. Activities and sources of exposure were significantly correlated with higher breath levels for the following chemicals:

benzene: visits to service stations, smoking, work in chemical and paint plants;
tetrachloroethylene: visits to dry cleaners.

6. Although questionnaires adequate for identifying household sources were not part of the study, the following sources were hypothesized:

p-dichlorobenzene: moth crystals, deodorizers, pesticides;
chloroform: hot showers, boiling water for meals;
styrene: plastics, insulation, carpets;
xylenes; ethylbenzene: paints, gasoline.

7. Residence near major outdoor point sources of pollution had little effect, if any, on personal exposure.

The TEAM direct approach has four basic elements:

- Use of a representative probability sample of the population under study
- Direct measurement of the pollutant concentrations reaching these people through all media (air, food, water, skin contact)
- Direct measurement of body burden to infer dosage
- Direct recording of each person's daily activities through diaries

The Denver - Washington, DC CO Exposure Study utilized a methodology for measuring the frequency distribution of CO exposures in a representative sample of urban populations during 1982-83.[20-22] Household data were collected from over 4400 households in Washington, DC and over 2100 households in the Denver metropolitan areas. Exposure data using personal monitors were collected from 814 individuals in Washington, DC, and 450 individuals in Denver, together with activity data from a stratified probability sample of the residents living in each of the two urban areas. Established survey sampling procedures were used. The resulting exposure data permit statistical comparisons between population subgroups (e.g., commuters vs. noncommuters, and residents with and without gas stoves). The data also provide evidence for judging the accuracy of exposure estimates calculated from fixed site monitoring data.

Additional efforts are underway to use these data to recognize indoor sources and factors which contribute to elevated CO exposure levels and to validate existing exposure models.

Microenvironment Models

Utilizing data collected in the Washington, DC urban-scale CO Study, two modeling and evaluation analyses have been developed. The first, conducted by Duan, is for the purpose of evaluating the use of microenvironmental and activity pattern data in estimating a defined population's exposure to CO.[16] The second, conducted by Flachsbart, is to model the microenvironmental situation of commuter rush-hour traffic (considering type and age of vehicle, speed, and meteorology) and observed CO concentrations.[5] With the assistance of a contractor, U.S. EPA has collected data on traffic variables, traffic volume, types of vehicles, and model year. An earlier study measured CO in a variety of microenvironments and under a variety of conditions.[23]

The indirect method for estimating population exposure to CO was compared to exposures to the CO concentrations observed while people carried personal monitors during their daily activities. The indirect estimate derived from personal monitoring at the low concentration levels, say 1 ppm but higher at levels above that. For example, at the 5 ppm level, indirect estimates were about half the direct estimates within the regression model utilizing these data. Although the results are limited, it appears that when monitoring experts design microenvironmental field surveys, there is a tendency to sample more heavily in those settings where the concentration is expected to be higher, thereby causing exaggerated levels of the indirect method. The possibility of using microenvironmental measurements and/or activity patterns from one city to extrapolate to those of another city is doubtful but not yet fully evaluated.

Dosimetry Research

The development of reliable biological indicators of either specific pollutant exposures or health effects is in its early stages. A limited number of biomarkers such as blood levels of lead or CO have been recognized and used for some time. Breath levels of VOCs or CO have also been measured successfully. However, the use of other biomarkers such as cotinine, a metabolite of nicotine, for a tracer compound of environmental tobacco smoke is still in its experimental phase. This also applies to use of the hydroxyproline-to-creatinine ratio as a measure of NO_2 exposure and also to use of DNA adducts which form as a result of VOC exposure and have been found to be correlated with genotoxic measures. Dosimetry methods development, though still very new and too often not yet ready for field application for humans, is obviously a very promising research area.

Exhaled breath measurements have been used successfuly in VOC and CO exposure studies. Since breath samples can be obtained noninvasively, they are preferred to blood measurements whenever they can meet the exposure research goals. A methodology to collect expired samples on a Tenax adsorbent has been developed and used on several hundred TEAM study subjects. Major findings have included the discovery that breath levels generally exceed outdoor levels, even in heavily industrialized petrochemical manufacturing areas. Significant correlations of breath levels with personal air exposures for certain chemicals give further proof that the source of the high exposure is in personal activities or indoors, at home as well as at work.

The basic advantages of monitoring breath rather than blood or tissues are:

1. Greater acceptability by volunteers. Persons give breath samples more readily than blood samples. The procedure is rapid and convenient, taking only 5-10 min. in all.

2. Greater sensitivity. Since volatile organic compounds often have a high air-to-blood partition coefficient, they will have higher concentrations in breath than in blood under equilibrium conditions. Thus, more than 100 compounds have been detected in the breath of subjects where simultaneously collected blood samples showed only one or two above detectable limits.

3. Fewer analytical problems. Several "clean-up" steps must be completed with blood samples, including centrifuging, extraction, etc., with each step carrying possibility for loss or contamination of the sample.

Measurements of CO in expired air often are used as indicators of carboxyhemoglobin (COHb) concentrations in blood, although the precise relationship between alveolar CO and blood COHb has not been agreed upon.

The U.S. EPA exposure monitoring program therefore included a breath monitoring component in its study of CO exposures in Denver and Washington, DC. The purpose was (1) to estimate the distribution of alveolar CO (and therefore blood COHb) concentrations in the nonsmoking adult residents of the two cities; and (2) to compare the alveolar CO measurements to preceding personal CO exposures.

The major findings of the breath monitoring program included:

1. The percent of nonsmoking adults with alveolar CO exceeding 10 ppm (i.e., blood COHb 2%) was 11% in Denver and 6% in Washington, DC.

2. The correlations between breath CO and previous 8-h CO exposure were 0.5 for Denver and 0.66 for Washington, DC.

3. The correlations between personal CO exposures at home or at work and ambient CO at the nearest stations averaged 0.25 at Denver and 0.19 at Washington, DC. Thus, the ambient data explained little of the variability of CO exposure.

Sampling Protocols

Statistical sampling protocols are the design for large-scale total human exposure field studies. They describe the procedures to be used in identifying respondents, choosing the sample sizes, selecting the number of persons to be contacted within various subpopulations, and other factors. They are essential to the total human exposure research program to ensure that a field survey will provide the information necessary to meet its objectives. Because one's activities affect one's exposures, another unique component of the total human exposure research program is the development of human activity pattern data bases. Such data bases provide a record describing what people do in time and space.

Whenever the objectives of a study are to make valid inferences beyond the group surveyed, a statistical survey design is required. For exposure studies, the only statistically valid procedure that is widely accepted for making such inferences is to select a probability sample from the target population. The survey designs used in the total exposure field studies have been three-stage probability-based, which consist of areas defined by census tracts, households randomly selected within the census tracts, and stratified sampling of screened eligible individuals.[20,24]

STATISTICAL ISSUES

TEAM Design Considerations

It appears that some variability in the TEAM exposure data might be due to meteorological factors such as some receptors being downwind of the sources while others are not. A more careful experimental design that includes consideration of these factors, including measurement of appropriate meteorological parameters, may lead to more meaningful data in future studies.

Other TEAM design considerations are:

1. The intraperson temporal variation in VOC exposure is crucial in risk assessment and should be given a high priority in future studies.

2. Given the substantial measurement error, the estimated exposure distributions can be substantially more heterogeneous than the true exposure distributions. For example, the variance of the estimated exposures is the sum of the variance of the true exposures and the variance of the measurement errors, assuming that: a) measurement errors are homoscedastic, and b) there is no correlation between measurement error and true exposure. Empirical Bayes methods are available for such adjustments.

3. The relatively high refusal rate in the sample enrollment is of concern. A more rigorous effort in the future to assess the impact of the refusal on the generalizability of the sample is desirable. For example, a subsample of the accessible part of the refusals can be offered an incentive to participate, or be offered a less intensive protocol for their participation; the data from the would-be refusals can then be compared with the "regular" participants to assess the possible magnitudes of selection bias.

4. In future studies, the following might be used:

   a. use of closed format questionnaires,
   b. use of artifical intelligence methodology,
   c. use of automated instrument output.

Development of Improved Microenvironmental Monitoring Designs

The direct method of personal exposure is appealing but is expensive and burdensome to human subjects. Monitoring microenvironments instead is less costly but estimtes personal exposure only indirectly. Obviously these approaches can be used in a complementary way to answer specific pollutant exposure questions.

With either method, a crucial issue is how to stratify the microenvironments into relatively homogeneous microenvironment types (METS).[12] Usually there are many possible ways to stratify the microenvironments into METs, thus there can be many potentially distinct METs. Obviously one cannot implement a stratification scheme with five hundred METs in field studies. It is therefore important to develop methods for identifying the most informative ways to stratify the microenvironments into METs. For example, if we can only afford to distinguish two METs in a field study, is it better to distinguish indoor and outdoor as the two METs, or is it better to distinguish awake and sleeping as the two METs?

Some of the more important issues which will require additional methodological development are:

1. How to identify the most informative ways to stratify microenvironments into METs.

2. How to optimize the number of METs, choosing between a larger number of METs and fewer microenvironments for each MET, and a smaller number of METs and more microenvironments for each MET.

3. How to allocate the number of monitored microenvironments across different METs: one should monitor more microenvironments for the more crucial METs (those in which the human subjects spend more of their time) than the less crucial METs.

Development and Validation of Improved Models for Estimating Personal Exposure from Microenvironmental Monitoring Data

Methodological development is needed for models which allow supplementing the direct personal monitoring approach with an activity diary enabling these data to be combined with indirect approach microenvironmental data to estimate personal exposure through a regression-like model. The basic exposure model which sums over microenvironments

$$E_i = \sum_j c_j t_{ij}$$

can be interpreted as a regression model with the concentrations being the parameters to be estimated. To fully develop this approach, it is necessary to make crucial assumptions about independence between individuals and between METs. Therefore, it is very important to validate the method empirically.

Errors-in-Variables Problem

It is important to recognize an errors-in-variables situation which may often occur in exposure assessment. In estimating the relationship between two variables, Y (a health effect) and X (true personal exposure), when X is not observed but a surrogate of X, say Z, which is related to X is observed. Such variables may have systematic errors as well as zero-centered random errors. The effects of the measurement bias are more serious in estimation situations than for hypothesis testing.

Choice Between Monitoring Instruments of Varying Precision and Cost

When designing monitoring programs, it is common to have available instruments of varying quality. Measurement devices that are less expensive to obtain and use are typically also less accurate and precise. Strategies could be developed and evaluated that consider the costs of measurement as well as the precision. In situations of high between-individual exposure variability, a less precise instrument of lower cost may be preferred if it permits an opportunity for enough additional study subjects.

Development of Designs Appropriate for Assessing National Levels

At the present time, the data available for the assessment of personal exposure distributions are restricted to a limited number of locales. The generalization from existing data to a very general population such as the national population requires a great deal of caution. However, it is conceivable that large scale studies or monitoring programs aimed at a nationally representative sample might be implemented in the future. It would be useful to consider the design of such studies using data presently available. It would also be useful to design studies of more limited scales to be conducted in the near future as pilot studies for a possible national study, so as to collect information which might be useful for the design of a national study.

An issue in the design of a national study is the amount of clustering of the sample: one has to decide how many locales to use, and how large a sample to take for each locale. The decision depends partly on the fixed cost in using additional locales, and partly on the intracluster correlation for the locales. For many of the VOC's measured in the TEAM studies, there is far more variability within locales than between locales, in other words, there is little intracluster correlation for the locales. This would indicate that a national study should be highly clustered, with a few locales and a large sample for each locale. On the other hand, if there is more variability between locales than within locales, a national study should use many locales and a small sample for each locale.

Further analysis of the existing TEAM data base can help to address these issues. For example, the TEAM sample to date can be identified as a "population" from which various "samples" can be taken. The characteristics of various sample types can be useful for the design of any followup studies as well as for a larger new study.

Evaluating Extreme Values in Exposure Monitoring

Short term extreme values of pollutant exposure may well be more important from a biological point of view than elevated temporal mean values. The study of statistical properties of extreme values from multivariate spatio-temporally dependent data is in its infancy. In particular, the possibility of synergy necessitates the development of a theory of multivariate extreme values. It is desirable to develop estimates of extreme quantiles of pollutant concentration.

Estimation Adjustment for Censored Monitoring Data

One should develop low exposure level extrapolation procedures and models, and check the sensitivity of these procedures to the models chosen. In some cases a substantial fraction of exposure monitoring data is below the detection limit even though these low exposure levels may be important. The problem of extrapolating from measured to unmeasured values thus naturally arises. Basically this is a problem of fitting the lower tail of the pollutant concentration distribution. Commonly used procedures assume either that below detectable level values are actually at the detection limit, or that they are zero, or that they are one-half of the detection limit.

In many monitoring situations we may find a good fit to simple models such as the lognormal for that part of the data which lies above the detection limit. Then the calculation of total exposure would use a lognormal extrapolation of the lower tail.

SUMMARY

Personal exposure assessment is a critical link in the overall risk assessment framework. Recent advances in exposure monitoring have provided new capabilities and additional challenges to the environmental research team, particularly to the statistician, to improve the current state of information on microenvironment concentrations, activity patterns, and particularly personal exposure. If these opportunities are realized, then risk assessments can more often use human exposure and risk data in addition to available animal toxicology information.

REFERENCES

1. Lioy, P. J., (1987) In Depth Exposure Assessments, JAPCA, 37, 791-793.

2. Epidemiology of Air Pollution, National Research Council National Academy Press, Washington, DC (1985), 1-334.

3. Ott, W. R. (1982) Concepts of human exposure to air pollution, Environ. Int., 7, 179-196.

4. Cortese, A. D. and Spengler, J.D. (1976) Ability of fixed monitoring stations to represent carbon monoxide exposure. J. Air Pollut. Control Assoc., 26, 1144.

5. Flachsbart, P. G. and Ott, W. R. (1984) Field Surveys of carbon monoxide in commercial settings using personal exposure monitors. EPA-600/4-94-019, PB-84-211291, U.S. Environmental Protection Agency, Washington, DC.

6. Wallace, L. A. (1979) Use of personal monitor to measure commuter exposure to carbon monoxide in vehicle passenger compartment. Paper No. 79-59.2, presented at the 72nd Annual Meeting of the Air Pollution Control Association, Cincinnati, OH.

7. Ott, W. R. and Eliassen, R. (1973) A survey technique for determining the representativeness of urban air monitoring stations with respect to carbon monoxide, J. Air. Pollut. Control Assoc. 23, 685-690.

8. Ott, W. R. and Flachsbart, P. (1982) Measurement of carbon monoxide concentrations in indoor and outdoor locations using personal exposure monitors, Environ. Int. 8, 295-304.

9. Peterson, W. B. and Allen, R. (1982) Carbon monoxide exposures to Los Angeles commuters, J. Air Pollut. Control Assoc. 32, 826-833.

10. Spengler, J. D. and Soczek, M. L. (1984) Evidence for improved ambient air quality and the need for personal exposure research, Environ. Sci. Technol. 18, 268-80A.

11. Ott, W. R. (1985) Total human exposure: An emerging science focuses on humans as receptors of environmental pollution, Environ. Sci. Technol. 19, 880-886.

12. Duan, N (1982) Models for human exposure to air pollutant, Environ. Int. 8, 305-309.

13. Mage, D. T. and Wallace, L. A., eds. (1979) Proceedings of the Symposium on the Development and Usage of Personal Monitors for Exposure and Health Effects Studies. EPA-600/9-79-032, PB-80-143-894, U.S. Environmental Protection Agency, Research Triangle Park, NC.

14. Wallace, L. A. (1981) Recent progress in developing and using personal monitors to measure human exposure to air pollution, Environ. Int. 5, 73-75.

15. Wallace, L. A. and Ott, W. R. (1982) Personal monitors: A state-of-the-art survey, J. Air Pollut. Control Associ. 32, 601-610.

16. Duan, N. (1984) Application of the microenvironment type approach to assess human exposure to carbon monoxide. Rand Corp., draft final report submitted to the U.S. Environmental Protection Agency, Research Triangle Park, NC.

17. Wallace, L. A., Zweidinger, R., Erickson, M., Cooper, S., Whitaker, D., and Pellizzari, E. D. (1982) Monitoring individual exposure: Measurements of volatile organic compounds in breathing-zone air, drinking water, and exhaled breath, Environ. Int. 8, 269-282.

18. Wallace, L., Pellizzari, E., Hartwell, T., Rosenzweig, M., Erickson, M., Sparacino, C. and Zelon, H. (1984) Personal exposures to volatile organic compounds: I. Direct measurements in breathing-zone air, drinking water, food, and exhaled breath, Environ. Res. 35, 293-319.

19. Wallace, L., Pellizzari, E., Hartwell, T., Zelon, H., Sparacino, C., and Whitmore, R. (1984) Analyses of exhaled breath of 335 urban residents for volatile organic compounds, in Indoor Air, vol. 4: Chemical Characterization and Personal Exposure, pp. 15-20. Swedish Council for Building Research, Stockholm.

20. Akland, G. G., Hartwell, T. D., Johnson, T.R., and Whitmore, R. W. (1985) Measuring human exposure to carbon monoxide in Washington, DC, and Denver, Colorado, during the winter of 1982-83, Environ. Sci. Technol. 19, 911-918.

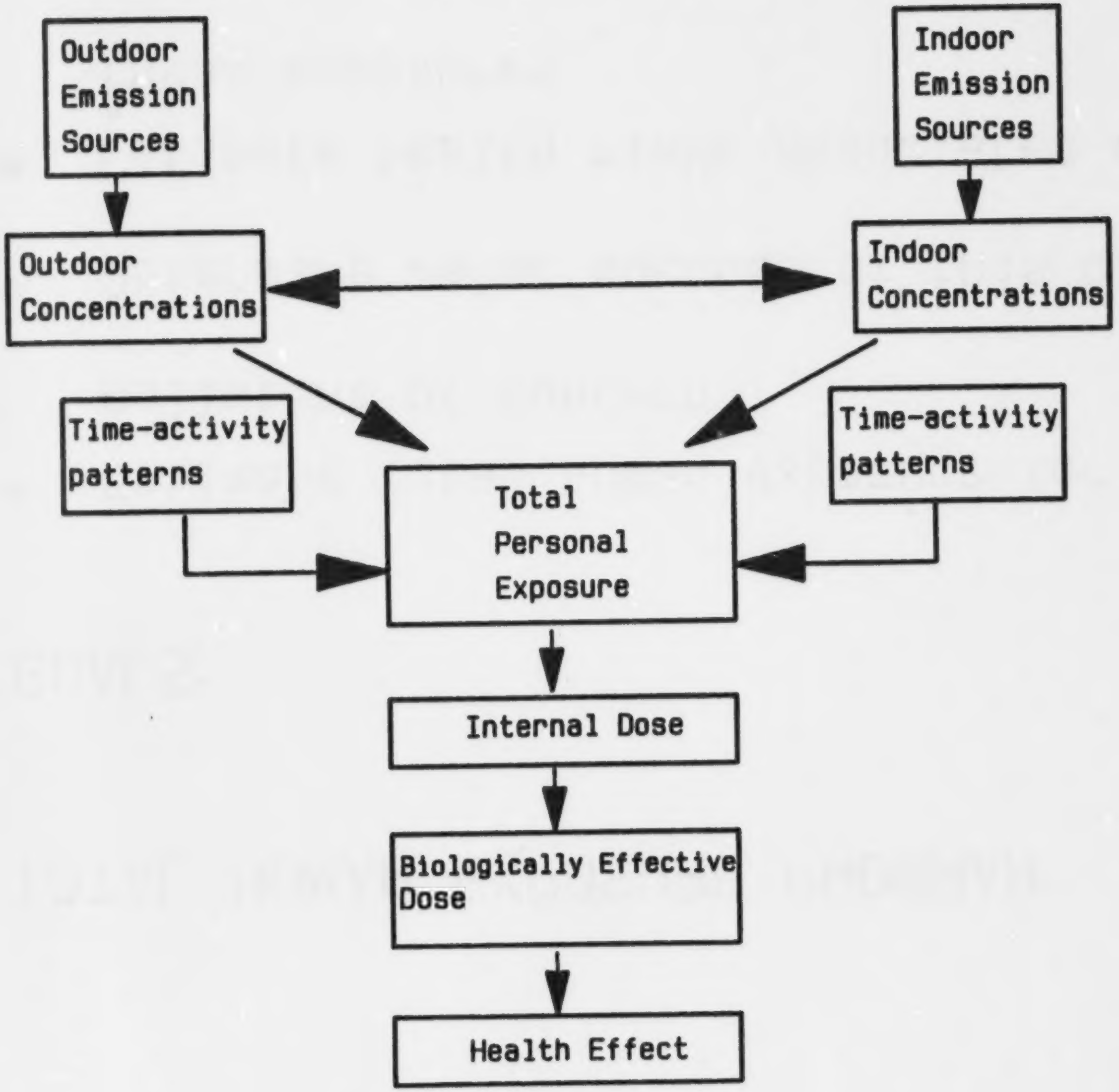
21. Johnson, T. (1984) A study of personal exposure to carbon monoxide in Denver, Colorado. EPA-600/4-84-015, PB-84-146-125, Environmental Monitoring Systems Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC

22. Hartwell, T. D., Carlisle, A. C., Michie, R. M., Jr., Whitmore, R. W., Zelon, H. S., and Whitehurst, D. A. (1984) A study of carbon monoxide exposure of the residents in Washington, DC. Paper No. 121.4, presented at the 77th Annual Meeting of the Air Pollution Control Association, San Francisco, CA.

23. Holland, D. M. and Mage, D. T. (1983) Carbon monoxide in four cities during the winter of 1981. EPA-600/4-83-025, Environmental Monitoring Systems Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC.

24. Whitmore, R. W., Jones, S. M., and Rozenzeig, M. S. (1984) Final sampling report for the study of personal CO exposure. EPA-600/S4-84-034, PB-84-181-957, Environmental Monitoring Systems Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC.

FRAMEWORK FOR EXPOSURE ASSESSMENT


Outdoor Emission Sources
Indoor Emission Sources
Outdoor Concentrations
Indoor Concentrations
Time-activity patterns
Total Personal Exposure
Time-activity patterns
Internal Dose
Biologically Effective Dose
Health Effect

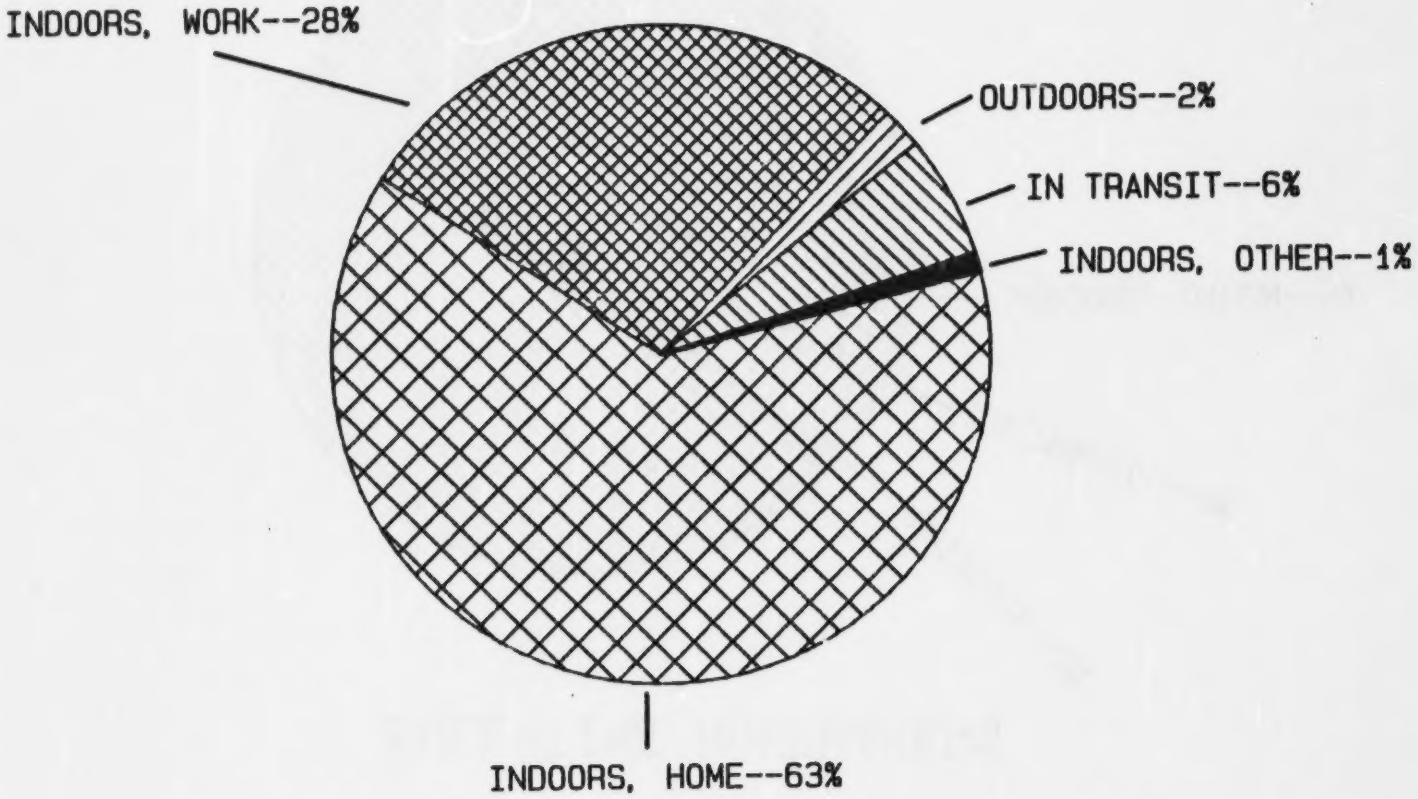
TOTAL HUMAN EXPOSURE PROGRAM

GOALS:

- Estimate total human exposure for each pollutant of concern
- Determine major sources of this exposure
- Estimate health risks associated with these exposures
- Determine actions to reduce these risks

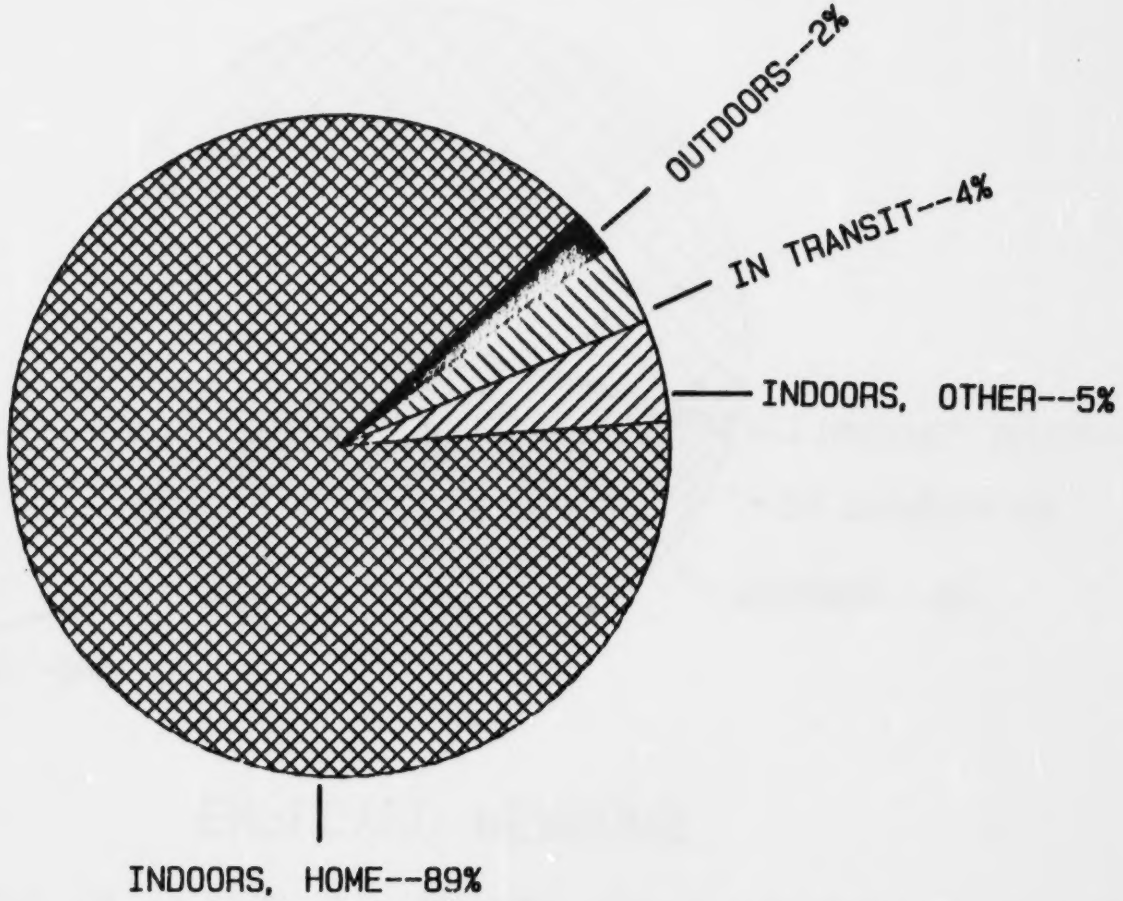
PROPORTION OF TIME IN SELECTED MICROENVIRONMENT

EMPLOYED PERSONS



PROPORTION OF TIME IN SELECTED MICROENVIRONMENTS

FULL-TIME HOMEMAKERS



MAJOR EXPOSURE SOURCES

Outdoors

Industrial
Automobile
Toxic wastes
Pesticides

Indoors

Tobacco smoke
Gas stoves
Cleaners
Sprays
Dry Cleaning
Paints
Polishes

EXPOSURE ASSESSMENT FOR COMMUNITY STUDIES

- Questionnaires
- Outdoor monitoring
- Indoor monitoring
- Personal monitoring
- Biological monitoring

DISCUSSION

William F. Hunt, Jr.
Chief, Monitoring and Research Branch
Technical Support Division
Research Triangle Park, NC 27711

William C. Nelson's paper provides an excellent overview of exposure monitoring and associated statistical issues. The reader must keep in mind that the paper is directed at estimating air pollution in microscale environments--in the home, at work, in automobiles, etc., as well as in the ambient air to which the general public has access.

While it is important to better understand air pollution levels in each of these microenvironments, it must be clearly understood that the principal focus of the nation's air pollution control program is directed at controlling ambient outdoor air pollution levels to which the general public has access. The Clean Air Act (CAA) of 1970 and the CAA of 1977 emphasized the importance of setting and periodically reviewing the National Ambient Air Quality Standards (NAAQS) for the nation's most pervasive ambient air pollutants--particulate matter, sulfur dioxide, carbon monoxide, nitrogen dioxide, ozone and lead. NAAQS(s) were set to protect against both public health and welfare effects.

One of these pollutants, carbon monoxide (CO), is discussed extensively in Dr. Nelson's paper. CO is a colorless, odorless, poisonous gas formed when carbon in fuels is not burned completely. Its major source is motor vehicle exhaust, which contributes more than two-thirds of all emissions nationwide. In cities or areas with heavy traffic congestion, however, automobile exhaust can cause as much as 95 percent of all emissions, and carbon monoxide concentrations can reach very high levels.

In Dr. Nelson's paper, he states that the correlations between personal CO exposures at home or at work and ambient CO at the nearest fixed site air monitoring stations are weak. This does not mean from an air pollution control standpoint, however, that there is something wrong with the fixed site CO monitoring network. As stated earlier, the air pollution control program is directed at controlling outdoor ambient air at locations to which the public has access. The microscale CO monitoring sites are generally located in areas of highest concentration within metropolitan areas at locations to which the general public has access.

The Federal Motor Vehicle Control Program has been very successful in reducing these concentrations over time. In fact, CO levels have dropped 32 percent between 1977 and 1986, as measured at the nation's fixed site monitoring networks.[1] This improvement has a corresponding benefit for people in office buildings which use the outdoor ambient air to introduce fresh air into their buildings through their ventilation systems. A major benefit occurs for people who are driving back and forth to work in their automobiles, for new cars are much less polluting than older cars. This should be clearly understood when trying to interpret the major findings of the breath monitoring programs that are described in Dr. Nelson's paper. Otherwise, the reader could mistakenly conclude that somehow the Federal Government may be in error in using fixed site monitoring. Such a conclusion would be incorrect. Further, it should be pointed out that a fixed site network also has the practical advantages of identifying the source of the problem and the amount of pollution control that would be needed.

Another area of concern that needs to be addressed in the future regarding the breath monitoring program is the relationship between alveolar CO and blood carboxyhemoglobin (COHb). Dr. Nelson states that the precise relationship between alveolar CO and blood COHb has not been agreed upon. Given that, is there an inconsistency in not being able to determine the relationship between alveolar CO and blood COHb and then using alveolar CO measurements in Washington, D.C. and Denver, Colorado to estimate blood COHb?

A final point, which needs to be addressed in the breath monitoring program,is the ability to detect volatile organic chemicals, some of which may be carcinogenic. What is the significance of being able to detect 100 compounds in breath, yet only one or two in blood above the detectable limits? Does the body expel the other 98 compounds that cannot be detected in the blood? If so, why?

STATISTICAL ISSUES

I agree with Dr. Nelson that meteorological factors should be incorporated into future TEAM studies, through more careful experimental design. The statistical issues identified under TEAM design considerations, the development of improved microenvironmental monitoring designs, errors-in-variables problem, choice between monitoring instruments of varying precision and cost, the development of designs appropriate for assessing

National levels, evaluating extreme values in exposure monitoring, and adjusting for censored monitoring data are all well thought out and timely. I strongly agree with his recommendation that when considering multiple pollutant species, as in the case of the volatile and semi-volatile organic chemicals, as well as polar compounds, the possibility of synergistic effects necessitates the development of a theory of multivariate extreme values.

SUMMARY

In conclusion, Dr. Nelson's paper provides a well thought out overview of exposure monitoring and the associated statistical issues. It should be an excellent reference for people interested in this topic. The reader should be aware, however, of the importance of the nation's fixed site monitoring network in evaluating the effectiveness of the nation's air pollution control program.

REFERENCE

1. _National Air Quality and Emissions Trends Report, 1986._ U.S. Environmental Protection Agency, Technical Support Division, Monitoring and Reports Branch, Research Triangle Park, NC 27711.

Designing Environmental Regulations

Søren Bisgaard and William G. Hunter*
Center for Quality and Productivity Improvement
University of Wisconsin-Madison
610 Walnut Street, Madison, Wisconsin 53705

■ Public debate on proposed environmental regulations often focuses almost entirely (and naively) on the allowable limit for a particular pollutant, with scant attention being paid to the statistical nature of environmental data and to the operational definition of compliance. As a consequence regulations may fail to accomplish their purpose. A unifying framework is therefore proposed that interrelates assessment of risk and determination of compliance. A central feature is the operating characteristic curve, which displays the discriminating power of a regulation. This framework can facilitate rational discussion among scientists, policymakers, and others concerned with environmental regulation.

Introduction

Over the past twenty years many new federal, state, and local regulations have resulted from heightened concern about the damage that we humans have done to the environment - and might do in the future. Public debate, unfortunately, has often focused almost exclusively on risk assessment and the allowable limit of a pollutant. Although this "limit part" of a regulation is important, a regulation also includes a "statistical part" that defines how compliance is to be determined; even though it is typically relegated to an appendix and thus may seem unimportant, it can have a profound effect on how the regulation performs.

Our purpose in this article is to introduce some new ideas concerning the general problem of designing environmental regulations, and, in particular, to consider the role of the "statistical part" of such regulations. As a vehicle for illustration, we use the environmental regulation of ambient ozone. Our intent is not to provide a definitive analysis of that particular problem. Indeed, that would require experts familiar with the generation, dispersion, measurements, and monitoring of ozone to analyze available data sets. Such detailed analysis would probably lead to the adoption of somewhat different statistical assumptions than we use. The methodology described below, however, can accommodate any reasonable statistical assumptions for ambient ozone. Moreover, this methodology can be used in the rational design of any environmental regulation to limit exposure to any pollutant.

Ambient Ozone Standard

For illustrative purposes, then, let us consider the ambient ozone standard (1,2). Ozone is a reactive form of oxygen that has serious health effects. Concentrations from about 0.15 parts per million (ppm), for example, affect respiratory mucous membranes and other lung tissues in sensitive individuals as well as healthy exercising persons. In 1971, based on the best scientific studies at the time, the Environmental Protection Agency (EPA) promulgated a National Primary and Secondary Ambient Air Quality Standard ruling that "an hourly average level of 0.08 parts per million (ppm) not to be" exceeded more than 1 hour per year." Section 109(d) of the Clean Air Act calls for a review every five years of the Primary National Ambient Air Quality Standards. In 1977 EPA announced that it was reviewing and updating the 1971 ozone standard. In preparing a new criteria document, EPA provided a number of opportunities for external review and comment. Two drafts of the document were made available for external review. EPA received more than 50 written responses to the first draft and approximately 20 to the second draft. The American Petroleum Institute (API), in particular, submitted extensive comments.

The criteria document was the subject of two meetings of the Subcommittee on Scientific Criteria for Photochemical Oxidants of EPA's Science Advisory Board. At each of these meetings, which were open to the public, critical review and new information were presented for EPA's consideration. The Agency was petitioned by the API and 29 member companies and by the City of Houston around the time the revision was announced. Among other things, the petition requested that EPA state the primary and secondary standards in such a way as to permit reliable assessment of compliance. In the Federal Register it is noted that

> EPA agrees that the *present deterministic form* of the oxidant standard has several limitations and has made reliable assessment of compliance difficult. *The revised ozone air quality standards are stated in a statistical form* that will more accurately reflect the air quality problems in various regions of the country and allow more reliable assessment of compliance with the standards. (Emphasis added)

Later, in the beginning of 1978, the EPA held a public meeting to receive comments from interested parties on the initial proposed revision of the standard. Here several representatives from the State and Territorial Air Pollution Program Administrators (STAPPA) and the Association of Local Air Pollution Control Officials participated. After the proposal was published in the spring of 1978, EPA held four public meetings to receive comments on the proposed standard revisions. In addition, 168 written comments were received during the formal comment period. The Federal Register summarizes the comments as follows:

> The majority of comments received (132 out of 168) opposed EPA's proposed standard revision, favoring either a more relaxed or a more

*) Deceased.

> stringent standard. State air pollution control agencies (and STAPPA) generally supported a standard level of 0.12 ppm on the basis of their assessment of an adequate margin of safety.
>
> Municipal groups generally supported a standard level of 0.12 ppm or higher, whereas most industrial groups supported a standard level of 0.15 ppm or higher. Environmental groups generally encouraged EPA to retain the 0.08 ppm standard.

As reflected in this statement, almost all of the public discussion of the ambient ozone standard (not just the 168 comments summarized here) focused on the limit part of the regulation. In this instance, in common with similar discussion of other environmental regulations, the statistical part of the regulation was largely ignored.

The final rule-making made the following three changes:

(1) The primary standard was raised to 0.12 ppm.

(2) The secondary standard was raised to 0.12 ppm.

(3) The definition of the point at which the standard is attained was changed to "when the expected number of days per calendar year" with maximum hourly average concentration above 0.12 ppm is equal to or less than one."

The Operating Characteristic Curve

Environmental regulations have a structure similar to that of statistical hypothesis tests. A regulation states how data are to be used to decide whether a particular site is in compliance with a specified standard, and a hypothesis test states how a particular set of data are to be used to decide whether they are in reasonable agreement with a specified hypothesis. Borrowing the terminology and methodology from hypothesis testing, we can say there are two types of errors that can be made because of the stochastic nature of environmental data: a site that is really in compliance can be declared out of compliance (type I error) and *vice versa* (type II error). Ideally the probability of committing both types of error should be zero. In practice, however, it is not feasible to obtain this ideal.

In the context of environmental regulations, an operating characteristic curve is the *probability of declaring a site to be in compliance* (d.i.c.) plotted as a function of some parameter θ such as the mean level of a pollutant. This $Prob\{d.i.c. \mid \theta\}$ can be used to determine the probabilities of committing type I and type II errors. As long as θ is below the stated standard, the probability of a type I error is $1 - Prob\{d.i.c. \mid \theta\}$. When θ is above the stated standard, $Prob\{d.i.c. \mid \theta\}$ is the probability of a type II error. Using the operating characteristic curve for the old and the new regulations for ambient ozone, we can evaluate them to see what was accomplished by the revision.

The old standard stated that "an hourly average level of 0.08 ppm [was] not to be exceeded more than 1 hour per year." This standard was therefore defined operationally in terms of the observations themselves. The new standard, on the other hand, states that the *expected number* of days per calendar year with a maximum hourly average concentration above 0.12 ppm should be less than one. Compliance, however, must be determined in terms of the *actual data*, not an unobserved *expected number*. How should this conversion be made? In Appendix D of the new ozone regulation, it is stated that:

> In general, the average number of exceedances per calendar year must be less than or equal to 1. In its simplest form, the number of exceedances at a monitoring site would be recorded for each calendar year and then averaged over the past 3 calendar years to determine if this average is less than or equal to 1.

Based on the stated requirements of compliance, we have computed the operating characteristic functions for the old and the new ozone regulations. They are plotted in Figures 1 and 2. (The last sentence in the legend for Figure 1 will be discussed below in the following section, Statistical Analysis.) To construct these curves, certain simplifying assumptions were made, which are discussed in the section entitled "Statistical Concepts." Before such curves are used in practice, these assumptions need to be investigated and probably modified.

According to the main part of the new ozone regulation, the interval from 0 to 1 expected number of exceedances of 0.12 ppm per year can be regarded as defining "being in compliance." Suppose the decision rule outlined above is used for a site that is operating at a level such that the expected number of days exceeding 0.12 ppm is just below one. In that case, as was noted by Javitz (3), with the new ozone regulation, there is a probability of approximately 37% in any given year that such a site will be declared out of compliance. Moreover, there is approximately a 10% chance of not detecting a violation of 2 expected days per year above the 0.12 ppm limit; that is, the standard operates such that the probability is 10% of not detecting occurrences when the actual value is twice its permissable value (2 instead of 1). Some individuals may find these probabilities (37% and 10%) to be surprisingly and unacceptably high, as we do. Others, however, may regard them as being reasonable or too low. In this paper, our point is not to pursue that particular debate. Rather, it is simply to argue that, before environmental regulations are put in place, different segments of society need to be aware of such operating characteristics, so that informed policy decisions can be made. It is important to realize that the relevant operating characteristic curves can be constructed *before* a regulation is promulgated.

Statistical Concepts

Let X denote a measurement from an instrument such that $X = \theta + \varepsilon$, where θ is the mean value of the pollutant and ε is the statistical error term with variance σ^2. The term ε contains not only the error arising from an imperfect instrument but also the fluctuations in the level of the pollutant itself. We assume that the measurement process is well calibrated and that the mean value of ε is zero. The *parameters* θ and σ^2 of the distribution of ε are unknown but *estimates* of them can be obtained from data. A prescription of how the data are to be collected is known as the *sampling plan*. It addresses the questions of how many, where, when, and how observations are to be collected. Any function $f(\underline{X}) = f(X_1, X_2, \ldots, X_n)$ of the observations is an *estimator*, for example, the average of a set of values or the number of observations in a sample above a certain limit. The value of the function f for a given sample is an *estimate*. The estimator has a distribution, which can be determined from the distribution of the observations and the functional form of the estimator. With the distribution of the estimator, one can answer questions of the form: what is the probability that the estimate $f = f(\underline{X})$ is smaller than or equal to some *critical value* c? Symbolically this probability can be written as $P = Prob\{f(\underline{X}) \leq c \mid \theta\}$.

If we want to have a regulation limiting the pollution to a certain level, it is not enough to state the limit as a particular value of a parameter. We must define compliance operationally in terms of the observations. The condition of compliance therefore takes the form of an estimator $f(X_1, \ldots, X_n)$ being less than or equal to some *critical value* c, that is, $\{f(X_1, \ldots, X_n) \leq c\}$. Regarded as a function of θ, the probability $Prob\{f(X_1, \ldots, X_n) \leq c \mid \theta\}$ is therefore the probability that the site will be declared to be in compliance with the regulation. It is, in fact, the operating characteristic function.

The operating characteristic function and consequently the probability of type I and type II errors are fixed by appropriate choice of the critical value and sampling plan. It is common statistical practice to specify a maximum type I error probability α and then to find a critical value c such that $Prob\{f(\underline{X}) \leq c \mid \theta_o\} = 1-\alpha$. To control the probability of type II errors, one would then design a sampling plan such that the probability of the type II error is at most β for a specified value θ_1 outside the compliance region. It is important to recognize that θ_0 and c are different; θ_0 is a point in the parameter space and c is a point in the sample space. Ignoring this subtle difference (which is almost always done in legal, legislative, and policymaking discussions) has led to unnecessary confusion. Because this difference exists, type I and type II errors exist. These errors should be confronted and balanced, not ignored.

Statistical Analysis

For purposes of illustration, let us consider the old and new regulations for ambient ozone. Let X denote the hourly average ozone level and let L be the limit, which for the *old* regulation was 0.08 ppm. Suppose the random variable X represents a single hourly average reading for ambient ozone that is independently and identically distributed. (This simplifying assumption is not necessary for application of this approach, but it is made here for X and below for Y for ease of exposition. Similar remarks apply to the assumptions of a normal distribution and a particular value of σ^2 stated below.) Denote by $p_L = Prob\{I_L(X) = 1\}$ the probability that X exceeds the limit L = 0.08 ppm. $I_L(x)$ is the indicator function, which is one for x>L and zero otherwise. A year consists of approximately $n = 365 \times 12 = 4380$ hours of observations (data are only taken from 9:01 am to 9:00 pm LST). The expected number of hours per year above the limit is then

$$\theta = E\{\sum_{i=1}^{4380} I_L(X_i) = 1\} = p_L \times 4380 .$$

The probability that a site is declared to be in compliance (d.i.c.) is

$$P_{old} = Prob\{\, d.i.c. \mid \theta\,\} = Prob\left\{\sum_{i=1}^{n} I(X_i) \leq 1 \mid \theta\right\} \qquad (1)$$

$$= \sum_{i=0}^{1} \binom{n}{i} p_L^i (1 - p_L)^{n-i} .$$

This probability P_{old}, plotted as a function of θ, is the operating characteristic curve for the old regulation (Figure 1). Note that *if* the old standard had been written in terms of an allowable limit of one for the *expected number* of exceedances above 0.08 ppm, the maximum type I error would be 1.00 − 0.73 = 0.27. The old standard, however, is actually written in terms of the *observed* number of exceedances so type I and type II errors, strictly speaking, are undefined.

The condition of compliance stated in the *new* regulation is that the "expected number of days per calendar year with daily maximum ozone" concentration exceeding 0.12 ppm must be less than or equal to 1." Let Y_j represent the daily maximum hourly average (j=1, . . . ,365). Suppose the random variables Y_j are independently and identically distributed. EPA proposed that the expected number of days (a parameter) be estimated by a three-year moving average of exceedances of 0.12 ppm. A site is in compliance when the moving average is less than or equal to 1. The expected number of days above the limit of L = 0.12 ppm is then

$$\theta = E\{\sum_{j=1}^{365} I_L(Y_j) = 1\} = 365 \times p_L .$$

The three-year specification of the new standard makes it hard to compare with the previous one-year standard. If, however, one computes the conditional probability that the number of exceedances in the present year is less than or equal to 0, 1, 2 and 3 and multiplies that by the probability that the number of exceedances was 3, 2, 1 and 0, respectively, for the previous two years, one then obtains a one-year operating characteristic function.

$$P_{new} = Prob\{\, d.i.c. \mid \theta\,\} = \sum_{k=0}^{3} Prob\,\{\, d.i.c. \mid k,\theta\,\}\, P(k)$$

where

$$P(k) = Prob\left\{\sum_{j=1}^{2\times 365} I(Y_j) = k\right\} = \binom{730}{k} p_L^k (1 - p_L)^{730-k} \quad (2)$$

and

$$Prob\{\, d.i.c. \mid k,\theta\,\} = \sum_{j=0}^{3-k} \binom{365}{j} p_L^j (1 - p_L)^{365-j} \quad (3)$$

where k=0,1,2,3. A plot of the operating characteristic function for the new regulation, P_{new} versus θ, is presented in Figure 2.

Figures 1 and 2 show the operating characteristic curves computed as a function of (1) the expected number of *hours* per year above 0.08 ppm for the old ambient ozone regulation and (2) the expected number of of *days* per year with a maximum hourly observation above 0.12 ppm for the new ambient ozone regulation. We observe that the 95 % *de facto* limit (the parameter value for which the site in a given year will be declared to be in compliance with 95 % probability) is 0.36 hours per year exceeding 0.08 ppm for the old standard and 0.46 days per year exceeding 0.12 ppm for the new standard. If the expected number of hours of exceedances of 0.08 ppm is one (and therefore in compliance), the probability is approximately 26% of declaring a site to be *not* in compliance with the old standard. If the expected number of days exceeding 0.12 ppm is one (and therefore in compliance), the probability is approximately 37% of declaring a site to be *not* in compliance with the new standard. (We are unaware of any other legal context in which type I errors of this magnitude would be considered reasonable.) Note that the parameter value for which the site in a given year will be declared to be in compliance with 95% probability is 0.36 hours per year exceeding 0.08 ppm for the old standard and 0.46 days per year exceeding 0.12 ppm for the new standard.

Neither curve provides sharp discrimination between "good" and "bad" values of θ. Note that the old standard did not specify any *parameter* value above which non-compliance was defined. The new standard, however, specifies that *one expected day* is the limit, thereby creating an inconsistency between what the regulation says and how it operates because of the large discrepancy between the stated limit and the operational limit.

The construction of Figures 1 and 2 only requires the assumption that the relevant observations are approximately identically and independently distributed (for the old standard, the relevant observations are those for the hourly ambient ozone measurements; for the new standard, they are the maximum hourly average measurements of the ambient ozone measurements each day). The construction does not require knowledge of the distribution of ambient ozone observations. If one has an estimate of this distributional form, however, a direct comparison of the new and old regulation is possible in terms of the concentration of ambient ozone (in units, say, of ppm.) To illustrate this point, suppose the random variable X_i is independently and identically distributed according to a normal distribution with mean μ and variance σ^2, that is, $X_i \sim N(\mu, \sigma^2)$. Then the probability of one observation being above the limit $L = 0.08$ is

$$Prob\{I(X) = 1\} = p_L = \Phi\left[\frac{\mu - L}{\sigma}\right] \tag{4}$$

where $\Phi(\ )$ is the cumulative density function of the standard normal distribution. The probability that a site is declared to be in compliance can be computed as a function of μ by substituting p_L from (4) into (1).

For the new regulation let X_{ij} represent the one-hour average, $(i=1, \ldots, 12; j=1, \ldots, 365)$, and $Y_j = \max\{X_{1j}, \ldots, X_{12,j}\}$. If $X_{ij} \sim N(\mu, \sigma^2)$, then $Y_j \sim H(y)$ where

$$H(y) = \left[\Phi\left[\frac{y-\mu}{\sigma}\right]\right]^{12} = Prob\{Y_j \le y\}$$

By substituting p_L in (2) and (3) with

$$p_L = Prob\{\,Y > 0.12\,\} = 1 - \left[\Phi\left[\frac{0.12 - \mu}{\sigma}\right]\right]^{12}$$

one obtains the operating characteristic function for the new standard.

For a fixed value of the variance σ^2, one can compute the operating characteristic curves for the old and new regulations to provide a graphical comparison of the way these two regulations perform. Figure 3 shows these curves for the old and new ambient ozone regulations computed as a function of the mean hourly values *when it is assumed that* $\sigma = 0.02$ *ppm*. We observe that the 95% de facto limit is changed from 0.0046 ppm to 0.045 ppm. That is, it is approximately ten times higher in the new ozone regulation.

We have three observations to offer with regard to the old and new regulations for ambient ozone standards. First, notwithstanding EPA's comment to the contrary, the new ozone regulation is not more statistical than the previous one; like all environmental regulations, both the new and old ozone regulations contain statistical parts, and, for that reason, both are statistical. Changing the specification from one in terms of a critical value to one in terms of a parameter does not make it more statistical. It actually introduced an inconsistency. The old standard did not specify any parameter value as a limit but only an operational limit in terms of the parameters. This therefore constitutes the standard. The new standard, however, specifies not only an intent in terms of what the desired limit is but also an operational limit. The large difference between the intented limit and the operational limit constitute the inconsistency. This inconsistency is a potential and unnecessary source of conflict. Second, the new regulation is dependent on the ambient ozone level for the past two years as well as the present year, which means that a sudden rise in the ozone level might be detected more slowly. The new regulation is also more complicated. Third, it is unwise first to record and store every single hourly observation and then to use only the binary observation as to whether the daily maximum is above or below 0.12 ppm. This procedure wastes valuable scientific information. As a matter of public policy, it is unwise to use the data in a binary form when they are already measured on a continuous scale. The estimate of the 1/365 percentile is an unreliable statistic. It is for this reason that type I and type II errors are as high as they are. In fact, the natural variability of this statistic is of the same order of magnitude as the change in the limit which was so much in debate.

If instead, for example, one used a procedure based on the t-statistic for control of the proportion above the limit, as is commonplace in industrial quality control procedures (4), one would get the operating characteristic curve plotted in Figure 4 (see also appendix). For comparison, the curve for the new regulation is also plotted as a function of the expected number of exceedances per year. With the new ozone regulation, the probability can exceed 1/3 that a particular site will be declared out of compliance when it is actually in compliance. The operating characteristic curve for the t-test is steeper (and hence has more discriminating power) than that for the new standard. The modified procedure based on the t-test generally reduces the probability that sites that are actually in compliance will be declared to be out of compliance. In fact, it is constructed so that there is 5% chance of declaring that a site is out of compliance when it is actually in compliance in the sense that the expected exceedance number is one per year. Furthermore, when a violation has occurred, it is much more certain that it will be detected with the t-based procedure. In this respect, the t-based procedure provides more protection to the public.

We do not conclude that procedures based on the t-test are best. We merely point out that there are alternatives to the procedures used in the old and new ozone standard. A basic principle is that information is lost when data are collected on a continuous scale and then reduced to a binary form. One of the advantages of procedures based on the t-test is that they do not waste information in this way.

The most important point to be made goes beyond the regulation of ambient ozone; it applies to regulation of all pollutants where there is a desire to limit exposure. With the aid of operating characteristic curves, informed judgements can be made when an environmental regulation is being developed. In particular, operating characteristic curves for alternative forms of a regulation can be constructed and compared *before* a final one is selected. Also, the robustness of a regulation to changes in assumptions, such as normality and statistical independence of observations, can be investigated prior to the promulgation. Note that environmental lawmaking, as it concerns the design of environmental regulations, is similar to design of scientific experiments. In both contexts, data should be collected in such a way that clear answers will emerge to questions of interest, and careful forethought can ensure that this desired result is achieved.

Scientific Framework

The operating characteristic curve is only one component in a more comprehensive scientific framework that we would like to promote for the design of environmental regulations. The key elements in this process are:

(a) Dose/risk curve
(b) Risk/benefit analysis
(c) Decision on maximum acceptable risk
(d) Stochastic nature of the pollution process
(e) Calibration of measuring instruments
(f) Sampling plan
(g) Decision function
(h) Distribution theory
(i) Operating characteristic function

Currently there may be some instances in which all of these elements are considered in some form when environmental regulations are designed. Because the particular purposes and techniques are not explicitly isolated and defined, however, the resulting regulations are not as clear nor as effective as they might otherwise be.

Often the first steps towards establishing an environmental regulation are (a) to estimate the relationship between the "dose" of a pollutant and some measure of health risk associated with it and (b) to carry out a formal or informal risk/benefit analysis. The problems associated with estimating dose/risk relationships and doing risk/benefit analyses are numerous and complex, and uncertainties can never be completely eliminated. As a next step a political decision is made - based on this uncertain scientific and economic groundwork - as to the maximum risk that is acceptable to society (c). As indicated in Figure 5, the maximum acceptable risk implies, through the dose/risk curve, the maximum allowable dose. The first three elements have received considerable attention when environmental regulations have been formulated, but the last six elements have not received the attention they deserve.

The maximum allowable dose defines the compliance set Θ_o and the noncompliance set Θ_1, which is its complement. The pollution process can be considered (d) as a stochastic process or statistical time-series $\phi(\theta; t)$. Fluctuations in the measurements X can usefully be thought of as arising from three sources: variation in the pollution level itself ϕ, the bias b in the readings, and the measurement error ε. Thus $X = \phi + b + \varepsilon$. Often it is assumed that $\phi = \theta$, a fixed constant and that variation arises only from the measurement error ε; however, all three components ϕ, b, and ε can vary. Ideally $b=0$ and the variance of ε is small.

Measurements will only have scientific meaning if there is a detailed operational description of how the measurements are to be obtained and the measurement process is in a state of statistical control. A regulation must include a specification relating to how the instruments are to be calibrated (e). These descriptions must be an integral part of a regulation if it is going to be meaningful. The subject of measurement is deeper than is generally recognized, with important implications for environmental regulation (5, 6, 7). The pollution process and the observed process as a function of time are indicated in Figure 5.

Logically the next question is (f) how best to obtain a sample $\underline{X} = (X_1, X_2, \ldots, X_n)$ from the pollution process. The answer to this question will be related to the form of the estimator $f(\underline{X})$ and (g) the decision rule.

$$d(f(\underline{X})) = \begin{cases} 0 : \textit{process in compliance} \\ 1 : \textit{process not in compliance} \end{cases}$$

The sample, the estimator, and the decision function are indicated in Figure 5. Based on knowledge about the statistical distribution of the sample (h), one can compute (i) the operating characteristic function $P = Prob\{d(f(\underline{X})) = 0 \mid \theta\}$ and plot the operating characteristic curve P versus θ. An operating characteristic function is drawn at the bottom of Figure 5. (In practice it would probably be desirable to construct more than one curve because, with different assumptions, different curves will result). Projected back on the dose/risk relationship (see Figure 5), this curve shows the probability of encountering various risks for different values of θ if the proposed environmental regulation is enacted. Suppose there is a reasonable probability that the pollutant levels occur in the range where the rate of change of the dose/risk relationship is appreciable; then the steeper the dose/risk function, the steeper the operating characteristic curve needs to be if the regulation is to offer adequate protection. The promulgated regulation should be expressed in terms of an operational definition that involves measured quantities, not parameters. Figure 5 provides a convenient summary of our proposed framework for designing environmental regulations.

In environmental lawmaking, it is most prudent to consider a range of plausible assumptions. Operating characteristic curves will sometimes change with different geographical areas to a significant degree. Although this is an awkward fact when a legislative, administrative, or other body is trying to enact regulations at an international, national, or other level, it is better to face the problem as honestly as possible and deal with it rather than pretending that it does not exist.

Operating Characteristic Curve as a Goal, Not a Consequence

We suggest that operating characteristic curves be published whenever an environmental regulation is promulgated that involves a pollutant the level of which is to be controlled. When a regulation is being developed, operating characteristic curves for various alternative forms of the regulation should be examined. An operating characteristic curve with specified desirable properties should be viewed as a goal, not as something to compute after a regulation has been promulgated. (Nevertheless, we note in passing that it would be informative to compute operating characteristic curves for existing environmental regulations.)

In summary, the following procedure might be feasible. First, based on scientific and economic studies of risks and benefits associated with exposure to a particular pollutant, a political decision would be reached concerning the compliance set in the form of an interval of the type $0 \leq \theta \leq \theta_o$ for a parameter of the distribution of the pollution process. Second, criteria for desirable sampling plans, estimators, and operating characteristic curves would be established. Third, attempts would be made to create a sampling plan and estimators that would meet these criteria. The costs associated with different sampling plans would be estimated. One possibility is that the desired properties of the operating characteristic curve might not be achievable at a reasonable cost. Some iteration and eventual compromise may be required among the stated criteria. Finally, the promulgated regulation would be expressed in terms of an operational definition that involves measured quantities, not parameters.

Injecting parameters into regulations, as was done in the new ozone standard, leads to unnecessary questions of interpretation and complications in enforcement. In fact, inconsistencies (such as that implied by $Prob\{f(\underline{X}) \leq c \mid \theta_0\} = 37\%$ for the new ozone standard) can arise when conceptual differences between c and θ_0 and between $f(\underline{X})$ and θ are ignored. These entities are commonly confused with one another and type I and type II errors are ignored. What is needed is a more refined conceptual model than that which underlies current environmental regulations, a model that makes these distinctions and acknowledges type I and type II errors.

Research Needs

Research that is used in designing environmental standards has focused on the first three elements of our framework (a), (b), and (c). If the last six elements do not receive relatively more attention than they currently receive, the precision obtained in estimating risk may well be lost by the lack of precision in estimating compliance. The above analysis, therefore, points to the need to have research resources more evenly spread among all the key elements (a), (b), ..., (i). Furthermore, more research needs to be conducted that takes a global view of how all the elements function together. It would be beneficial to analyze many of the already promulgated standards using the framework outlined above and in particular to compute operating characteristic curves. Such research will sometimes require the development of new distribution theory because standards typically use rather complex decision rules. Moreover, most environmental data are serially correlated and consequently the shape of the operating characteristic function will be affected. At present little statistical theory is developed to cope with this problem. Preliminary studies we have done show that operating characteristic curves for binary sampling plans as used in the ozone standard seem to be seriously affected by serial correlation. Monte Carlo simulation might prove a viable alternative to distribution theory in evaluating the operating characteristic function for complex decision rules and serially correlated time series.

In our discussion above we only considered one pollutant and its regulation. The interaction among several pollutants and other environmental factors, however, might create higher risks than would be anticipated from separate studies on the individual pollutants themselves. Such issues are only beginning to be addressed (8).

A related issue is the problem of what constitutes a rational attitude towards risk. It seems irrational to impose strict standards for one pollutant when other equally hazardous pollutants have much more relaxed standards. A harmonization among standards seems desirable. In order to address such issues it is necessary to develop methods for comparing convolutions of probability of occurrence, dose/risk relationships, and operating characteristic functions for several pollutants simultaneously. This will require an extension of the framework outlined above to multiple pollutants. However, that framework can be used as a first step in attacking these more comprehensive problems that are so important to protecting our environment.

Conclusion

One of the purposes of environmental law, which has been defined as the rules for planetary housekeeping (9), is to prevent harm to society. Assessment of risk is one of the key issues in environmental lawmaking and continued research is needed on how to measure risk and make decisions regarding risk; but risk assessment is not enough. If laws with good operating characteristics are not designed, the effort expended on risk assessment will simply be wasted. With limited resources, we need to develop methods for economically and rationally allocating resources to provide high levels of safety. Ideally a system of environmental management and control should be composed of individual laws that limit potential risk in a consistent manner. The ideas outlined in this article give partial answers to two connected questions: (i) how can we formulate an *individual* quantitative regulation so that it will be scientifically sound and (ii) how can we construct a rational *system* of environmental regulations?

If the framework outlined above is used properly in the course of developing environmental regulations, some of the important operating properties of different alternatives would be known. The public would know the probilities of violations not being detected (type II errors); industries would know the probabilities of being accused incorrectly of violating standards (type I errors); and all parties would know the costs associated with various proposed environmental control schemes. We believe that the operating characteristic curve is a simple, yet comprehensive device for presenting and comparing different alternative regulations because it brings into the open many relevant and sometimes subtle points. For many people it is unsettling to realize that type I and type II errors will be made, but it is unrealistic to develop regulations pretending that such errors do not occur. In fact, one of the central issues that should be faced in formulating effective and fair regulations is the estimation and balancing of the probabilities of such occurrences.

Acknowledgments

This research was supported by grants SES - 8018418 and DMS - 8420968 from the National Science Foundation. Computing was facilitated by access to the research computer at the Department of Statistics, University of Wisconsin, Madison.

Appendix

The t-statistic procedure is based on the estimator $f(\underline{x}) = (L-\bar{x})/s$ where L is the limit (0.12 ppm), $\bar{x}$ the sample average, and s the sample standard deviation. The decision function is

$$d(f(\underline{x})) = \begin{cases} f(\underline{x}) \geq c : \textit{in compliance} \\ f(\underline{x}) < c : \textit{not in compliance} \end{cases} \qquad \text{(A1)}$$

The critical value c is found from the requirement that

$$Prob\left\{ \frac{L-\bar{x}}{s} > c \,\middle|\, \frac{L-\mu}{\sigma} = z_0 \right\} = 1-\alpha \qquad \text{(A2)}$$

where $z_0 = \Phi^{-1}(1-\theta_0)$ and θ_0 is the fraction above the limit we at most want to accept (here 1/365).

The exact operating characteristic function is found by reference to a non-central t-distribution, but for all practical purposes the following approximation is sufficient:

$$Prob\left\{ \frac{L-\bar{x}}{s} > c \,\middle|\, \theta \right\} = \Phi\left[\frac{\sqrt{n}\,(\Phi^{-1}(1-\theta) - c)}{\sqrt{1+c^2/2}} \right] . \qquad \text{(A3)}$$

The operating characteristic function in Figure 4 is constructed using $\alpha=0.05$, $\theta_0=1/365$ and $n=3\times365$. Substituting (A3) into (A2) yields

$$\Phi\left[\frac{\sqrt{n}\,(\Phi^{-1}(1-\theta_0) - c)}{\sqrt{1+c^2/2}} \right] = 1 - 0.05 \qquad \text{(A4)}$$

which solved for the critical value yields $c = 2.6715$. Refer for example to (4) for more details.

Literature Cited

(1) National Primary and Secondary Ambient Air Quality Standards, *Federal Register 36*, **1971** pp 8186-8187. (This final rulemaking document is referred to in this article as the *old* ambient ozone standard.)

(2) National Primary and Secondary Ambient Air Quality Standards, *Federal Register 44*, **1979** pp 8202-8229. (This final rulemaking document is referred to in this article as the *new* ambient ozone standard.) The background material we summarize is contained in this comprehensive reference.

(3) Javitz, H. J. *J. Air Poll. Con. Assoc.* **1980** 30, pp 58-59.

(4) Hald, A. "Statistical Theory with Engineering Applications"; Wiley, New York, 1952; pp 303-311.

(5) Hunter, J. S. *Science 210*, **1980** pp 869-874;

(6) Hunter, J. S. In "Appendix D", *Environmental Monitoring*, Vol IV, National Academy of Sciences 1977;

(7) Eisenhart, C. In "Precision Measurements and Calibration", National Bureau of Standards Special Publication 300 Vol. 1, 1969; pp 21-47.

(8) Porter, W. P.; Hinsdill, R.; Fairbrother, A.; Olson, L. J.; Jaeger, J.; Yuill, T.; Bisgaard, S.; Hunter, W. G.; K. Nolan, K. *Science* **1984**, 224, pp 1014-1017.

(9) Rogers, W. H. *"Handbook of Environmental Law"*; West Publishing Company, 1977, St. Paul, MN.

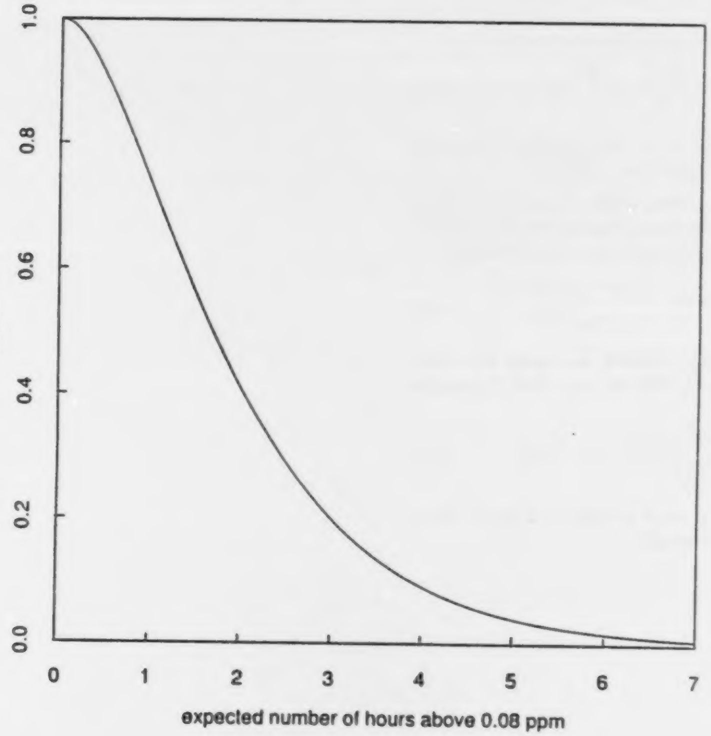
Figure 1. Operating characteristic curve for the 1971 ambient ozone standard (old standard), as a function of the expected number of hours of exceedances of 0.08 ppm per year. Note that *if* the old standard had been written in terms of an allowable limit of one for the ***expected number*** of exceedances above 0.08 ppm, the maximum type I error would be 1.00 - 0.73 = 0.27.

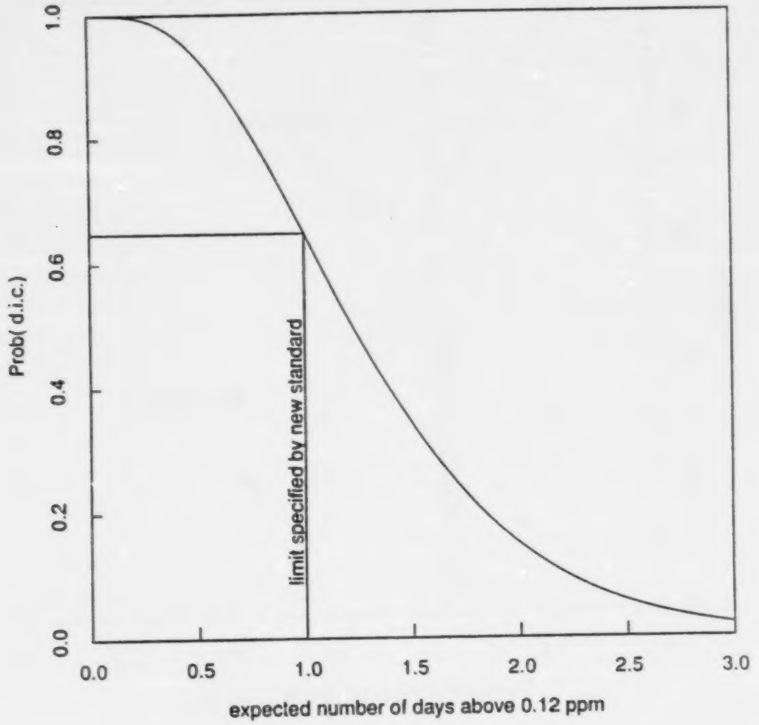
Figure 2. Operating characteristic curve for the 1979 ambient ozone standard (new standard), as a function of the expected number of days of exceedances of 0.12 ppm per year. Note that the maximum type I error is 1.00 – 0.63 = 0.37.

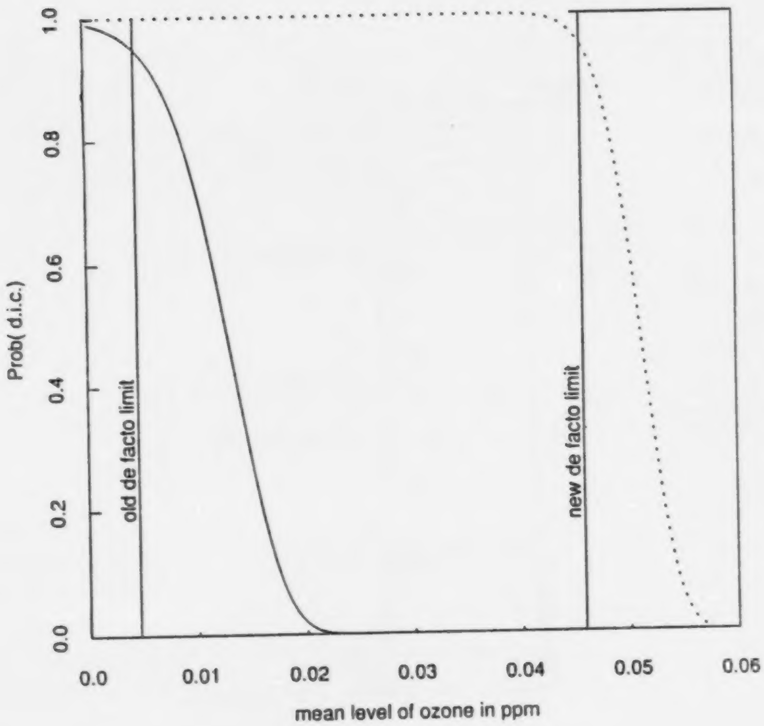
Figure 3. Operating characteristic curves for the old and the new standards as a function of the mean value of ozone measured in parts per million when it is assumed that ozone measurements are normally and independently distributed with $\sigma = 0.02$ ppm.

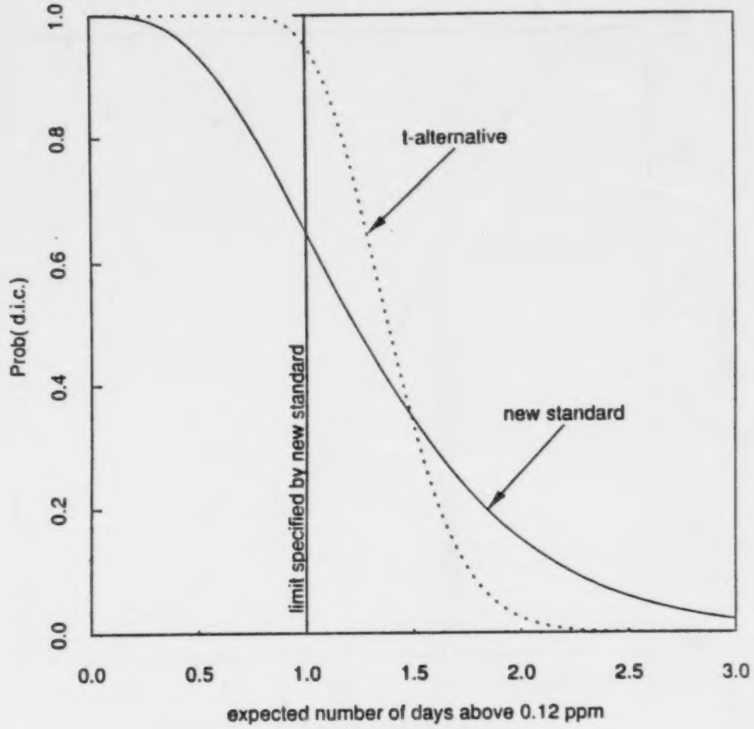
Figure 4. Operating characteristic curves for the new ozone standard and a t-statistic alternative as a function of the expected number of exceedances per year.

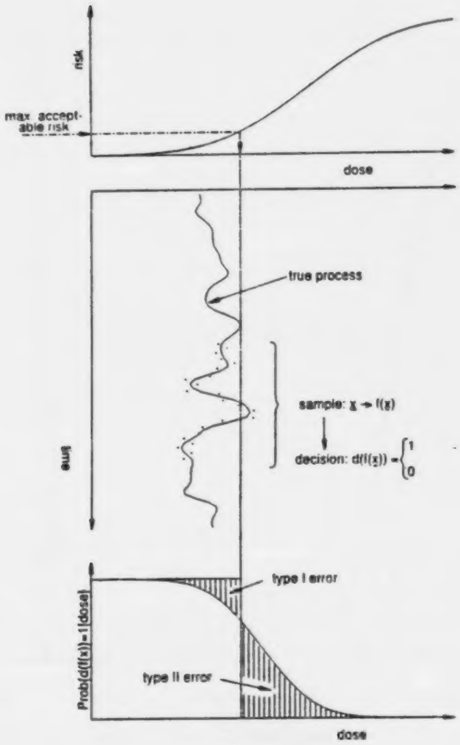
Figure 5. Elements of the environmental standard-setting process: Laboratory experiments and/or epidemiological studies are used to assess the dose/risk relationship. A maximum acceptable risk is determined through a political process balancing risk and economic factors. The maximum acceptable risk implies a limit for the "dose" which again implies a limit for the pollution process as a function of time. Compliance with the standard is operationally determined based on a discrete sample $\underline{x}$ taken from a particular site. The decision about whether a site is in compliance is reached through use of a statistic f and a decision function d. Knowing the statistical nature of the pollution process, the sampling plan, and the functional form of the statistics and the decision function, one can compute the operating characteristic function. Projecting the operating characteristic function back on the dose/risk relationship, one can assess the probability of encountering various levels of undetected violation of the standard.




Prob(d.i.c.)
0.0
0.2
0.4
0.6
0.8
1.0
limit specified by new standard
0.0
0.5
1.0
1.5
2.0
2.5
3.0
expected number of days above 0.12 ppm


Prob(d.i.c.)
0.0
0.2
0.4
0.6
0.8
1.0
old de facto limit
new de facto limit
0.0
0.01
0.02
0.03
0.04
0.05
0.06
mean level of ozone in ppm


1.0
0.8
0.6
0.4
0.2
0.0
Prob(d.i.c.)
t-alternative
new standard
limit specified by new standard
0.0
0.5
1.0
1.5
2.0
2.5
3.0
expected number of days above 0.12 ppm


risk
max. acceptable risk
dose
true process
sample: $\underline{x} \rightarrow f(\underline{x})$
decision: $d(f(\underline{x})) = \begin{cases} 1 \\ 0 \end{cases}$
time
type I error
type II error
Prob[d(f(x))=1|dose]
dose

DISCUSSION

W. Barnes Johnson

EPA PROGRAMS AND ENVIRONMENTAL STANDARDS

I appreciate the general points that Dr. Bisgaard has made regarding the development of environmental standards. I agree that generally, when standards are developed, most of the technical emphasis is placed on developing the magnitude of the absolute number, which Dr. Bisgaard calls the "limit part" of the standard. In contrast, frequently little work is expended developing the sampling program and the rules that are used to evaluate compliance with the limit in application, which he calls the "statistical part" of the standard. At EPA some programs do a thorough and thoughtful job of designing environmental standards. However, other EPA programs could benefit from Dr. Bisgaard's work because they have focused strictly on the magnitude of the standard and have not considered the "statistical part" of the standard.

However, I insist that the ozone standard and all of the National Ambient Air Quality Standards fall into the category of standards where both the "limit part" and the "statistical part" of the standard have been designed based on extensive performance evaluations and practical considerations.

There are other EPA programs that have also done an excellent job of designing and evaluating the "limit part" and the "statistical part" of their standards. For example, under the Toxic Substances Control Act (TSCA) regulations, there are procedures for managing PCB containing wastes. In particular, PCB soil contamination must be cleaned up to 50 ppm. Guidances have been prepared that stipulate a detailed sampling and evaluation program and effectively describe the procedure for verifying when the 50 ppm limit has been achieved. Also under the TSCA mandate, clearance tests are under development for verifying that, after the removal of asbestos from a building, levels are not different from background levels.

There are, however, many programs at EPA that have not performed the analysis and inquiry necessary to design the "statistical part" of their standards. One example is the Maximum Contaminant Levels (MCLs) which are developed and used by EPA'S drinking water program. MCLs are concentration limits established for controlling pollutants in drinking water supplies. Extensive health effect, engineering, and economic analysis is used to choose the MCL concentration value. However, relatively little work is done to ensure that, when compliance with the MCL is evaluated, appropriate sampling and analysis methodologies are used to ensure a designed level of statistical performance.

Similarly, risk-based cleanup standards are used in EPA's Superfund program as targets for how much abandoned hazardous waste sites should be cleaned up. These are concentration levels either borrowed from another program (e.g., an MCL) or developed based on site-specific circumstances. A great deal of effort has been expended on discussions of how protective the actual risk related cleanup standards should be; however, virtually no effort has been focused on the methodology that will be used to evaluate attainment of these standards. Drinking water MCLs and Superfund cleanup standards could benefit from the approaches offered by Dr. Bisgaard.

PRACTICAL ENVIRONMENTAL STANDARDS DESIGN: POLITICS, POLLUTANT BEHAVIOR, SAMPLING AND OBJECTIVES

Dr. Bisgaard clearly points out that his use of the ozone standard is only for the purpose of example and that the message of his presentation applies to the development of any standard. I have responded by trying to identify other EPA program areas that could benefit from the perspective offered by Dr. Bisgaard's approach. However, it is important to realize that the development of the "statistical part" of an environmental standard must consider the nature of the political situation, pollutant behavior, sampling constraints, and the objective of the standard. Ignorance of these practical considerations can limit the usefulness of a proposed standard regardless of the theoretical basis. The developers of the ozone standard were quite aware of these contingencies and it is reflected in the form of the "statistical part" of the ozone standard.

Central Tendency Versus Extremes

I must agree that a standard based on central tendency statistics will be more robust with better operating characteristics than a standard based on peak statistics. The difficulty is that EPA is not concerned with estimating or controlling the mean ozone concentration. Ozone is a pollutant with acute health effects and, as such, EPA's interest lies in control of the extremes of the population. Peak statistics were the primary concern when the ozone standard was developed.

EPA, in the development of NAAQS's, has tried to balance statistical performance with objectives by examining the use of other statistics that are more robust and yet retain control of the extremes. For example, EPA has suggested basing the standard on the fourth or fifth largest value; however, commenters maintained that EPA would lose control of the extremes and cause undo harm to human health. It has also been suggested that the peak to mean ratio (P/M) be considered. The problem with this approach is that the P/M is highly variable across the United States because of variation in the "ozone season." The objective of developing a nationally applicable regulatory framework would be quite difficult if each locale was subject to a different standard.

Decision Errors and Power

In addition, regardless of the standard that is chosen, decision errors will be highest when the true situation at a monitoring station is at or close to the standard. As the true situation becomes well above or below the standard, certainty increases and our decisions become less subject to error. Of course, it would be most desirable to have an operating characteristic function with a large distinct step at the standard. This operating characteristic would have no error even when the true situation is slightly above or below the standard; however, this is virtually impossible. Therefore, when standards are compared for their efficacy, it is important to compare performance along the continuum when the true situation is well above, at, and well below the standard. One should not restrict performance evaluation to the area at or immmediately adjacent to the standard, for most statistics the performance will be quite low in this region.

Dr. Bisgaard points out from his Figure 2 that when a site is in compliance and at the standard, expecting to exceed the standard on one day, there is a 37% chance that the site may be indicated as exceeding the standard. However, it can also be shown that when a site is below the standard and expects to exceed the standard on one-half of a day, there is only about a 6% chance that the site may be indicated as exceeding the standard. Conversely, it can be pointed out that when the site is above the standard and expects to exceed the standard on three days, there is only a 3% chance that the site will be found to be in compliance.

Dr. Bisgaard is quite correct in pointing out that the operating characteristics of a standard based on the mean are better than a standard based on the largest order statistic. However, as mentioned above, a standard based on the mean does not satisfy the objectives of the ozone standard. EPA staff have tendered proposals to improve the operating characteristics of the standard. One of these involved the development of a three-tiered approach that would allow a site to be judged: in attainment, not in attainment, or too close to call. The existing structure of the attainment program was not flexible enough to permit this approach.

Pollutant Behavior

Ozone is a pollutant which exists in the environment at a high mean ambient level of approximately one-third the existing standard. Effort expended trying to drive down peak statistics indirectly by controlling the mean would be futile. This is because mean levels can only be reduced to the background mean which, relative to the standard, is high even in the absence of air pollution.

Another point to consider is that ozone behavior is influenced by both annual and seasonal meteorological effects. This is the reason that the newest standard is based on three years of data. The effect of an extreme year is reduced by the averaging process associated with a three year standard. As mentioned above, work has also focused on controlling the peak to mean ratios; however, because ozone seasons vary radically across the country, this sort of measure would be difficult to implement.

Dr. Bisgaard has also questioned the new standard because of the use of the term "expected." This terminology was probably included in the wording because of the many legal and policy edits that are performed on a draft regulation. It was not intended that the term "expected" be applied in the technical statistical use of the term. The term was intended to show that EPA had considered and reflected annual differences in ozone conditions in the three year form of the standard.

CONCLUSIONS

Dr. Bisgaard brings an interesting and useful perspective to the development of environmental standards. The important idea is that an environmental standard is more than a numerical limit and must include a discussion of the associated sampling approach and decision function. I tried to extend this central idea by adding two primary points. First, there are several programs within EPA that can benefit from Dr. Bisgaard's perspective; however, the NAAQS program is fully aware of and has considered these sampling and decision issues in exhaustive detail. Second, the practical issues that influence the implementation of an environmental standard are a primary constraint and must be understood in order to develop a standard that offers a useful measure of compliance.

QUALITY CONTROL ISSUES IN TESTING COMPLIANCE WITH A REGULATORY STANDARD: CONTROLLING STATISTICAL DECISION ERROR RATES

by

Bertram Price
Price Associates, Inc.

prepared under

EPA Contract No. 68-02-4139
Research Triangle Institute

for

The Quality Assurance Management Staff
Office of Research and Development
U. S. Environmental Protection Agency
Washington, D.C. 20460

ABSTRACT

Testing compliance with a regulatory standard intended to control chemical or biological contamination is inherently a statistical decision problem. Measurements used in compliance tests exhibit statistical variation resulting from random factors that affect sampling and laboratory analysis. Since a variety of laboratories with potentially different performance characteristics produce data used in compliance tests, a regulatory agency must be concerned about uniformity in compliance decisions. Compliance monitoring programs must be designed to avoid, for example, situations where a sample analyzed by one qualified laboratory leads to a noncompliance decision, but there is reasonable likelihood that if the same sample were analyzed by another qualified laboratory, the decision would be reversed.

Two general approaches to designing compliance tests are discussed. Both approaches have, as an objective, controlling statistical decision error rates associated with the compliance test. One approach, the approach typically employed, depends on interlaboratory quality control (QC) data. The alternative, referred to as the intralaboratory approach, is based on a protocol which leads to unique QC data requirements in each laboratory. An overview of the statistical issues affecting the development and implementation of the two approaches is presented and the approaches are compared from a regulatory management perspective.

SECTION 1 - INTRODUCTION

Testing compliance with a regulatory standard intended to control chemical or biological contamination is inherently a statistical decision problem. Measurements used in compliance tests exhibit statistical variation resulting from random factors affecting sampling and laboratory analysis. Compliance decision errors may be identified with Type I and Type II statistical errors (i.e., false positive and false negative compliance test results, respectively). A regulating agency can exercise control over the compliance testing process by establishing statistical decision error rate objectives (i.e., error rates not to be exceeded). From a statistical design perspective, these error rate objectives are used to determine the number and types of measurements required in the compliance test.

Bias and variability in measurement data are critical factors in determining if a proposed compliance test satisfies error rate objectives. Various quality control (QC) data collection activities lead to estimates of bias and variability. An interlaboratory study is the standard approach to obtaining these estimates. (The U.S. Environmental Protection Agency [USEPA] has employed the interlaboratory study approach extensively to establish bias and variability criteria for test procedures required for filing applications for National Pollution Discharge Elimination System [NPDES] permits - 40 CFR Part 136, Guidelines Establishing Test Procedures for the Analysis of Pollutants Under the Clean Water Act.) An alternative means of estimating bias and variability that does not require an interlaboratory study is referred to in this report as the intralaboratory approach. The intralaboratory approach relies on data similar to those generated in standard laboratory QC activities to extract the information on bias and variability needed for controlling compliance test error rates.

The purpose of this report is to describe and compare the interlaboratory and intralaboratory approaches to collecting QC data needed for bias and variability estimates which are used in compliance tests. Toward that end, two statistical models, which reflect two different attitudes toward compliance test development, are introduced. Model 1, which treats differences among laboratories as random effects, is appropriate when the laboratory producing the measurements in a particular situation is not uniquely identified, but is viewed as a randomly selected choice from among all qualified laboratories. If Model 1 is used, an interlaboratory study is necessary to estimate "between laboratory" variance which is an essential component of the compliance test. Model 2 treats laboratory differences as fixed effects (i.e., not random, but systematic and identified with specific laboratories). If Model 2 is used, bias adjustments and estimates of variability required for compliance tests are prepared in each laboratory from QC data collected in the laboratory. Model 2 does not require estimates of bias and variability from interlaboratory data.

The remainder of this report consists of five sections. First, in Section 2, statistical models selected to represent the data used in compliance tests are described. In Section 3, a statistical test used in compliance decisions is developed. The comparison of interlaboratory and intralaboratory approaches is developed in two steps. Section 4 is included primarily for purposes of exposition. The types and numbers of measurements needed for a compliance test are derived assuming that the critical variance components - i.e., within and between laboratories - have known values. This section provides the structure for comparing the interlaboratory and intralaboratory approaches in the realistic situation where the variance components must be estimated. The comparison is developed in Section 5. A summary and conclusions are presented in Section 6.

SECTION 2 - STATISTICAL MODELS

Compliance tests are often complex rules defined as combinations of measurements that exceed a quantitative standard. However, a simple rule - an average of measurements compared to the standard - is the basis for most tests. This rule provides the necessary structure for developing and evaluating the interlaboratory and intralaboratory approaches. Throughout the subsequent discussion, the compliance standard is denoted by C_0 and interpreted as a concentration - e.g., micrograms per liter. Samples of the target medium are obtained, analyzed by chemical or other appropriate methods and summarized as an average for use in the test. The statistical design issues are:

- total number of measurements required;
- number and type of samples required; and
- number of replicate analyses per sample required.

The design issues are resolved by imposing requirements on the compliance test error rates (i.e., the Type I and Type II statistical error rates).

Many sources of variation potentially affect the data used in a compliance test. The list includes variation due to sample selection, laboratory, day and time of analysis, analytical instrument, analyst, and measurement error. To simplify the ensuing discussion, the sources have been limited to sample selection, laboratory, and measurement error. (Measurement error means analytical replication error or single analyst variability.) This simplification, limiting the number of variance components considered, does not limit the generality of subsequent results.

The distribution of the compliance data is assumed to have both mean and variance proportional to the true concentration. (This characterization has been used since many types of environmental measurements reflect these properties.) The data, after transformation to logarithms, base e, may be described as:

EQ 1 $Y_{i,j,k} = \mu + B_i + S_{i,j} + \epsilon_{i,j,k}$

where i = 1(1)I refers to laboratory, j = 1(1)J refers to sample and k = 1(1)K refers to analytical replication. Two different interpretations referred to as Model 1 and Model 2 are considered for the factors on the right side of equation 1.

In Model 1:

μ - ln(C), where C is the true concentration;

B_i - the logarithm of recovery (i.e., the proportion of the true concentration recovered by the analytical method) which is a laboratory specific effect treated as random with mean zero and variance σ^2_B;

$S_{i,j}$ - a sample effect which is random with mean zero and variance σ^2_S; and

$\epsilon_{i,j,k}$ - replication error which is random with mean zero and variance σ^2_ϵ.

It follows that:

$$E[Y_{i,j,k}] = \mu$$

$$Var[Y_{i,j,k}] = \sigma^2_B + \sigma^2_S + \sigma^2_\epsilon$$

and denoting as $\bar{Y}_i$ an average over samples and replicates,

EQ 2 $Var[\bar{Y}_i] = \sigma^2_B + \sigma^2_S/J + \sigma^2_\epsilon/J \cdot K.$

In Model 2, B_i is interpreted as a fixed effect (i.e., B_i is bias associated with laboratory i). All other factors have the same interpretation used in Model 1. Therefore, in Model 2:

$$E[Y_{i,j,k}] = \mu + B_i$$

$$\mathrm{Var}[Y_{i,j,k}] = \sigma^2{}_S + \sigma^2{}_\epsilon$$

and

EQ 3 $$\mathrm{Var}[\bar{Y}_i] = \sigma^2{}_S/J + \sigma^2{}_\epsilon/J\cdot K$$

Differentiating between Model 1 and Model 2 has significant practical implications for establishing an approach to compliance testing. These implications are developed in detail below. For now, it is sufficient to note that the collection of B_i's are treated as scaler factors uniquely associated with laboratories. If the identity of the specific laboratory conducting an analysis is unknown because it is viewed as randomly selected from the population of all laboratories, then B_i is treated as a random effect. If the laboratory conducting the analysis is known, B_i is treated as a scaler, namely the bias of the ith laboratory.

SECTION 3 - STATISTICAL TEST: GENERAL FORMULATION

The statistical test for compliance is based on an average of measurements, Y. Assuming that Y's are normally distributed (recall that Y is the natural logarithm of the measurement), noncompliance is inferred when

EQ 4 $$\bar{Y} > T$$

where T and the number of measurements used in the average are determined by specifying probabilities of various outcomes of the test. (For simplicity in exposition in this section, the subscripts i, j, and k used to describe the models in Section 2 are suppressed. Also, σ_Y is used in place of the expressions in EQ 2 and EQ 3 to represent the standard deviation of Y. The more detailed notation of EQ 2 and EQ 3 is used in the subsequent sections where needed.)

Let p_1 and p_2 be probabilities of declaring noncompliance when the true means are $d_1 \cdot C_0$ and $d_2 \cdot C_0$ respectively $(d_1, d_2 > 0)$, and let

$$\mu_0 = \ln(C_0)$$

$$D_1 = \ln(d_1),\ D_2 = \ln(d_2).$$

Requiring

EQ 5 $\quad p_1 = P[\ \bar{Y} > T:\ \mu = \mu_0 + D_1\ ]$

and

EQ 6 $\quad p_2 = P[\ \bar{Y} > T:\ \mu = \mu_0 + D_2\ ]$

leads to values of T and the number of measurements used to form Y by solving

EQ 7 $\quad [(T - \mu_0 + D_1)]/\sigma_{\bar{Y}} = Z_{1-p1}$

and

EQ 8 $\quad [(T - \mu_0 + D_2)]/\sigma_{\bar{Y}} = Z_{1-p2}$

where Z_{1-p1} and Z_{1-p2} are percentile points of the standard normal distribution.

The solutions are:

EQ 9 $\quad T = \sigma_{\bar{Y}} \cdot Z1_{-p1} + \mu_0 + D_1$

EQ 10 $\quad \sigma_{\bar{Y}} = (D_2 - D_1)/(Z_{1-p1} - Z_{1-p2}).$

This formulation allows considerable flexibility for determining compliance test objectives. Consider the following three special cases:

Case (i). When $d_1 = 1$, $p_1 = \alpha$, d_2 is any positive number greater than 1 and $p_2 = 1 - \beta$, the formulation reduces to the classical hypothesis testing problem H_0: $\mu = \mu_0$ versus H_1: $\mu = \mu_0 + D_2$. The correct number of measurements establishes the probabilities of Type I and Type II errors at α and β respectively.

Case (ii). Let $d_1 = 1$, d_2 be a positive number less than 1, $p_1 = 1 - \beta$, and $p_2 = \alpha$. This formulation also reduces to the classical hypothesis testing problem H_0: $\mu = \mu_0 + D_2$ versus H_1: $\mu = \mu_0$. (Note that $\mu_0 + D_2 < \mu_0$, i.e., $D_2 < 0$.)

Case (iii). Let $1 < d_1 < d_2$. Set $p_1 < p_2$ to large values (e.g., .90 and .99). This formulation imposes a high probability of failing the compliance test when the mean is D_1 times the standard, and a higher probability of failing when the mean is further above the standard.

Case (ii) imposes a more stringent regulatory program on the regulated community than Case (i). In Case (i), the regulated community may establish control methods to hold the average pollution level at the standard. In Case (ii), the pollution level must be controlled at a concentration below the standard if the specified error rates are to be achieved. In Case (iii), a formal Type I error is not defined. Individual members of the regulated community may establish the Type I error rate by setting their own pollution control level - the lower the control level, the lower the Type I error rate. In Case (iii), the regulated community has another option also. There is a tradeoff between the control level and the number of measurements used in the compliance test. Individuals may choose to operate at a level near the standard and increase the number of measurements used in the compliance test over the number required to achieve the stated probability objectives. The important difference between Case (iii) and the two other cases is the responsibility placed with the regulated community regarding false alarms (i.e., Type I errors). Since false alarms affect those regulated more than the regulator, Case (iii) may be the most equitable approach to compliance test formulation.

SECTION 4 - SAMPLE SIZE REQUIREMENTS: VALUES OF VARIANCE COMPONENTS KNOWN

The discussion below follows the structure of Case (i) described above. Based on the general formulation developed in Section 3, the conclusions obtained also hold for Cases (ii) and (iii).

MODEL 1

The compliance test is a statistical test of:

$$H_O: \quad \mu = \mu_O = \ln(C_O)$$

versus

$$H_1: \quad \mu = \mu_O + D_2$$

where C_O is the compliance standard. Assuming the values of the variance components are known, the test statistic is

$$Z = (\bar{Y}_i - \mu_O)/(\sigma^2{}_B + \sigma^2{}_S/J + \sigma^2{}_\epsilon/J\cdot K)^{1/2}.$$

Specifying the Type I error rate to be α leads to a test that rejects H_O if

EQ 11 $$Z > Z_{1-\alpha}$$

where $Z_{1-\alpha}$ is the $(1-\alpha)$th percentile point of the standard normal distribution. If the Type II error is specified to be β when the alternative mean is $\mu_O + D_2$, then:

EQ 12 $$\sigma^2{}_B + \sigma^2{}_S/J + \sigma^2{}_\epsilon/J\cdot K = [D_2/(Z_{1-\alpha} - Z_{1-\beta})]^2.$$

Any combination of J and K satisfying EQ 12 will achieve the compliance test error rate objectives. However, unique values of J and K may be determined by minimizing the cost of the data collection program subject to the constraint in EQ 12. Total cost may be stated as:

EQ 13 $$TC = J \cdot C_1 + J \cdot K \cdot C_2$$

where C1 is the unit cost of obtaining a sample and C_2 is the cost of one analysis.

Using the LaGrange Multiplier method to minimize EQ 13 subject to the constraint imposed by EQ 12 yields:

EQ 14 $$K = (\sigma_\epsilon/\sigma_S) \cdot (C_1/C_2)^{1/2}$$

and

EQ 15 $$J = [\sigma_S \cdot \sigma_\epsilon/(U-\sigma^2{}_B)] \cdot [\sigma_S/\sigma_\epsilon + (C_2/C_1)^{1/2}]$$

where

$$U = [D_2/(Z_{1-\alpha} + Z_{1-\beta})]^2.$$

(If EQ 14 does not produce an integer value for K, the next largest integer is used and J is adjusted accordingly.)

The number of replicate analyses for each sample, K, increases as the ratio of the sampling cost to the analysis cost increases and the ratio of the single analyst standard deviation to the sampling standard deviation increases. In many situations, the analysis cost, C_2, is much larger than the sampling cost, C_1, and the sampling variance is much larger than single analysis variability. Under these conditions, the number of replicate analyses, K, will be 1 (i.e., each sample will be analyzed only once).

MODEL 2

Since

$$E(\bar{Y}_i) = \mu + B_i$$

the statistic used in the compliance test must incorporate a bias adjustment (i.e., an estimate of B_i). This can be achieved by analyzing standard samples prepared with a known concentration C. (Choosing C at or near C_0 minimizes the effects of potential model specification errors.) Let

EQ 16 $$b_{i,j,k} = Y_{i,j,k} - \ln C = B_i + S'_{i,j} + \epsilon_{i,j,k}.$$

Since

$$E(\bar{b}_i) = B_i$$

b_i is an estimate of B_i and

$$\mathrm{Var}(\bar{b}_i) = \sigma^2_{S'}/J' + \sigma^2_{\epsilon}/J'\cdot K'$$

where

$S'_{i,j}$ - an effect associated with standard samples which is random with mean zero and variance $\sigma^2_{S'}$;

J' - the number of standard samples used to estimate B_i; and

K' - the number of analyses conducted on each standard sample.

(Note that single analyst variability, σ^2_{ϵ}, is assumed to have the same value for field samples and prepared samples.)

The test statistic is

EQ 17 $$(\bar{Y}_i - b_i - \mu_0)/[\sigma^2_S/J + \sigma^2_{S'}/J' + \sigma^2_{\epsilon}/(1/J'\cdot K' + 1/J\cdot K)]^{1/2}$$

The cost function used to allocate the samples and replicates is:

EQ 18 $$TC = J \cdot C_1 + J' \cdot C_3 + (J \cdot K + J' \cdot K') \cdot C_2$$

where C_3 is the unit cost for preparing a standard sample. Type I and Type II error rates - α and β - are achieved if:

EQ 19 $$\sigma^2{}_S/J + \sigma^2{}_{S'}/J' + \sigma^2{}_\epsilon(1/J' \cdot K' + 1/J \cdot K) = U$$

where

$$U = [D_2/(Z_{1-\alpha} + Z_{1-\beta})]^2,$$

as defined in the discussion of Model 1.

Minimizing costs subject to the constraint on variance yields

EQ 20 $$K = (\sigma_\epsilon/\sigma_S) \cdot (C_1/C_2)^{1/2},$$

which is identical to the solution obtained for Model 1, and

EQ 21 $$K' = (\sigma_\epsilon/\sigma_{S'}) \cdot (C_3/C_2)^{1/2},$$

EQ 22 $$J' = (\sigma_{S'}/U) \cdot [\sigma_S \cdot (C_1/C_3)^{1/2}, + 2 \cdot \sigma_\epsilon(C_2/C_3)^{1/2} + \sigma_{S'}],$$

and

EQ 23 $$J = J' \cdot (\sigma_S/\sigma_{S'}) \cdot (C_3/C_1)^{1/2}.$$

The solutions for K and K' are similar. Each increases with the ratio of sampling to analytical costs and the ratio of analytical to sampling standard deviations.

SECTION 5 - SAMPLE SIZE REQUIREMENTS: VALUES OF VARIANCE COMPONENTS UNKNOWN

In this section the interlaboratory and intralaboratory approaches for obtaining estimates of the variance components necessary to implement the designs developed in Section 4 are described. As in Section 4, the design objective is to control the compliance test error rates (i.e., the Type I and Type II error probabilities). The discussion is simplified by considering situations where the cost of analysis is significantly greater than the cost of sampling, and the sample to sample variability is at least as large as the analytical variability:

$$C_2 >> C_1 \text{ and } \sigma^2_S > \sigma^2_\epsilon.$$

Under these conditions, K = 1 (i.e., each sample is analyzed only once). Also, the value of K' determined from EQ 21 (i.e., the number of replicate analyses performed on each standard sample), will be set equal to 1 since the cost of preparing standard samples for estimating B_i is significantly less than the cost of analyzing those samples (i.e., $C_3 << C_2$).

When K = K' = 1, the variances used to define the test statistic are, for Model 1 and Model 2 respectively:

EQ 24 $$\mathrm{Var}(\bar{Y}_i) = \sigma^2_B + (\sigma^2_S + \sigma^2_\epsilon)/J$$

$$= \sigma^2_B + \sigma^2_{\epsilon'}/J$$

and

EQ 25 $$\mathrm{Var}(\bar{Y}_i - \bar{b}_i) = (\sigma^2_S + \sigma^2_\epsilon)/J + (\sigma^2_{S'} + \sigma^2_\epsilon)/J'$$

$$= \sigma^2_{\epsilon'}/J + \sigma^2_{\epsilon''}/J'.$$

(The notations $\sigma^2_{\epsilon'}$ and $\sigma^2_{\epsilon''}$ reflect the addition of the two variances indicated in Equations 24 and 25.)

MODEL 1

A compliance test designed on the basis of Model 1 requires estimates of $\sigma^2_{\epsilon'}$ and σ^2_B. An estimate of σ^2_B can be obtained only from an interlaboratory study. $\sigma^2_{\epsilon'}$ also may be estimated using interlaboratory data or it may be estimated from the J measurements of field samples used to form the average when the compliance test is performed.

As described by Youden (1975), an interlaboratory study involves M laboratories (between 6 and 12 are used in practice) which by assumption under Model 1 are randomly selected from the collection of all laboratories intending to produce measurements for compliance testing. For the discussion below, let n denote the number of samples analyzed by each laboratory. (Youden recommends n = 6 prepared as 3 pairs where the concentrations of paired samples are close to each other but not identical.)

Let

$$W_{i,j} = \ln(V_{i,j}/C_j)$$

where $\{V_{i,j}: i=1(1)M; j=1(1)n\}$ are the measurements produced by the i-th laboratory on the j-th sample, and $\{C_j: j=1(1)n\}$ are the concentration levels used in the study. (Youden does not recommend using logarithms, however the logarithmic transformation is convenient and is consistent with other assumptions in Youden's design.) The statistical model describing the interlaboratory study measurements is:

EQ 26 $$W_{i,j} = B_i + \epsilon''_{i,j}$$

where

B_i is an effect associated with the i-th laboratory and treated as a random variable with mean zero and variance σ^2_B; and

$\epsilon''_{i,j}$ is analytical error, the sum of single analyst error and an effect associated with variation among standard samples, which has mean zero and variance $\sigma^2_{\epsilon''}$.

Using standard ANOVA (analysis of variance) techniques, σ^2_B may be estimated from the "within laboratory" and "between laboratory" mean squares, Q_1 and Q_2:

EQ 27 $\quad Q_1 = \Sigma(W_{i,j} - \bar{W}_i)^2/M\cdot(n-1)$

and

EQ 28 $\quad Q_2 = n\cdot\Sigma(\bar{W}_i - \bar{W})^2/(M-1)$.

The estimate is:

EQ 29 $\quad s^2_B = (Q_2 - Q_1)/n$

which reflects differences among the laboratories through the quantity

EQ 30 $\quad \Sigma(B_i - \bar{B})^2$.

Also, Q_1 is an estimate of $\sigma^2_{\epsilon''}$.

The compliance test statistic may be defined either as

EQ 31a $\quad R = (\bar{Y}_i - \mu_O)/(s^2_B + Q_1/J)^{1/2}$

or

EQ 31b $\quad R = (\bar{Y}_i - \mu_O)/(s^2_B + S^2_{\epsilon'}/J)^{1/2}$

where $s^2_{\epsilon'}$ is the sample variance of the J measurements,

$$s^2_{\epsilon'} = \Sigma(Y_{i,j} - \bar{Y}_i)^2/(J - 1)$$

and $\{Y_{i,j} = \ln(X_{i,j}),\ j = i(i)J\}$ are the measurements obtained from field samples in the laboratory selected to conduct the analyses. (Based on the discussion at the beginning of this section, K is always equal to 1. Therefore, the notation describing compliance measurements has been simplified, i.e., $Y_{i,j} \equiv Y_{i,j,1}$). Note that Q_1 estimates the average variability over laboratories, whereas $s^2_{\epsilon'}$ estimates variability for the laboratory conducting the test. Also, Q_1 is an estimate of $\sigma^2_{\epsilon''}$, the variability associated with the analysis of standard samples; $s^2_{\epsilon'}$ is an estimate of the variability associated with the analysis of field samples.

The ratios in EQ 31a and EQ 31b have approximate t-distributions when the null hypothesis is true. The degrees of freedom may be estimated by methods developed by Satterthwaite (1946). Although it is possible to approximate the degrees of freedom and use a percentile point of the t-distribution to define the test, that approach is complicated. Develop it at this point would be an unnecessary diversion. Instead, non-compliance will be inferred when

EQ 32 $\quad R > Z_{1-\alpha}$

where $Z_{1-\alpha}$ is the $(1 - \alpha)$th percentile point of the standard normal distribution. (If R has only a few degrees of freedom, which is likely, the Type I error rate will be larger than α. The situation may be improved by using, for example, $Z_{1-\alpha/2}$ or some other value of Z larger than $Z_{1-\alpha}$. If necessary, exact values of Z could be determined using Monte Carlo methods.)

The number of samples, J, that must be analyzed for the compliance test is obtained by specifying that the expression in EQ 32 is equal to $1-\beta$ when the true mean is $\mu_0 + D_2$. The value of J may be obtained either by using approximations based on the normal distribution, the noncentral t-distribution, or by estimates based on a Monte Carlo simulation of the exact distribution of R.

If EQ 31a is used, the compliance test criterion (i.e., the expression in EQ 32) becomes

EQ 33 $$GM(X_{i,j}) > C_O \cdot \exp[Z_{1-\alpha} \cdot (s^2{}_B + Q_1/J)^{1/2}]$$

where GM is the geometric mean of the J compliance measurements. The right side of the inequality is a fixed number once the interlaboratory study is completed. The advantage of this approach is the simplicity realized in describing the compliance test to the regulated community in terms of one measured quantity, the geometric mean. The disadvantage is using Q_1 rather than the sample variance calculated from the compliance test measurements which is likely to be a better estimate of variability for the particular laboratory conducting the test.

MODEL 2

Under Model 2, estimates of variance from interlaboratory study data are unnecessary. Since the laboratory conducting the analyses for the compliance test is uniquely identified, the laboratory factor, B_i, is a scaler, and the variance component, $\sigma^2{}_B$, does not enter the model. The variance estimates needed for the compliance test can be obtained from the measurements used to compute $\overline{Y}_i$ and $\overline{b}_i$.

The test statistic is

EQ 34 $$t = (\overline{Y}_i - \overline{b}_i - \mu_O)/(s^2{}_{\epsilon'}/J + s^2{}_{\epsilon''}/J')^{1/2}$$

which has an approximate t-distribution with degrees of freedom equal to J + J' - 2 when the true mean is μ_O. (The statistic would have an exact t-distribution if $\sigma^2{}_{\epsilon'}$ were equal to $\sigma^2{}_{\epsilon''}$.) Noncompliance is inferred if

EQ 35 $t > t_{1-\alpha}$.

J and J' are determined by requiring that the probability of the expression in EQ 35 be equal to 1 - β when the true mean is $\mu_O + D_2$. This calculation can be made using the noncentral t-distribution. Where $\sigma^2_{\epsilon'} = \sigma^2_{\epsilon''}$, the noncentrality parameter is $D_2/[\sigma^2_\epsilon \cdot (1/J + 1/J')]$. (Note that this formulation implies a tradeoff between J and J' for achieving the compliance test error rate objectives.) If $\sigma^2_{\epsilon'}$ and $\sigma^2_{\epsilon''}$ are not equal, the correct value to replace $t_{1-\alpha}$ in EQ 35 and values of J and J' may be determined using Monte Carlo methods.

SECTION 6 - DISCUSSION AND CONCLUSIONS

Both statistical models considered above are consistent with reasonable approaches to compliance testing. The two approaches, however, have distinctly different data requirements.

Model 1, through EQ 32a, reflects "the conventional" approach to compliance testing. A "target value for control," C_O, is established (e.g., either a health based standard or a "best available control technology" standard) and then adjusted upward to account for both analytical variability and laboratory differences. Using EQ 33, noncompliance is inferred when the geometric mean of the compliance test measurements, $GM(X_{i,j})$, is larger than C_O multiplied by a factor which combines estimates reflecting variability between laboratories, σ^2_B, and analytical variability within laboratories. Since an estimate of σ^2_B is required in the Model 1 approach, an interlaboratory study is required also. The role of σ^2_B, which reflects laboratory differences, is to provide insurance against potentially conflicting compliance results if one set of samples were analyzed in two different laboratories. Systematic laboratory differences (i.e., laboratory bias) could lead to a decision of noncompliance based on analyses conducted in one laboratory and a decision of compliance based on analyses of the same samples conducted in another laboratory.

In practice, σ^2_B is replaced by s^2_B, an estimate obtained from the interlaboratory study. The variability of this estimate also affects the compliance test error rates. If the variance of s^2_B is large, controlling the compliance test error rates becomes complicated. Requiring that more field samples be analyzed (i.e., increasing J) may help. However, increasing the amount of interlaboratory QC data to reduce the variance of s^2_B directly may be the only effective option. Based on interlaboratory QC data involving 6 to 12 laboratories, which is current practice, the error in s^2_B as an estimate of σ^2_B is likely to be as large as 100%. If interlaboratory QC data were obtained from 30 laboratories, the estimation error still would exceed 50%. (These results are based on a 95% confidence interval for σ^2_B/s^2_B determined using the chi-square distribution.) Since interlaboratory data collection involving 12 laboratories is expensive and time consuming, it is doubtful if a much larger effort would be feasible or could be justified.

Using Model 2 and the intralaboratory approach, a regulatory agency would not attempt to control potential compliance decision errors resulting from laboratory differences by using an estimate of "between laboratory" variability to adjust the compliance standard. Instead, compliance data collected in each laboratory would be adjusted to reflect the laboratory's unique bias and variability characteristics. In many situations, bias for any specific laboratory can be estimated as precisely as needed using QC samples. Also, the variance of the bias estimate, which is needed for the compliance test, can be estimated from the same set of QC sample measurements. An estimate of analytical variability required for the compliance test can be estimated from the measurements generated on field samples. Therefore, all information needed to develop the compliance test can be obtained within the laboratory that produces the measurements for the test.

From a regulatory management perspective, both approaches (i.e., Model 1 using interlaboratory QC data and Model 2 using intralaboratory QC data) lead to compliance tests that satisfy specified decision error rate objectives. However, the intralaboratory approach based on Model 2 appears to be the more direct approach. The design for producing data that satisfy error rate objectives is laboratory specific, acknowledging directly that laboratories not only have different bias factors, but also may have different "within laboratory" variances. Each laboratory estimates a bias adjustment factor and a variance unique to that laboratory. Then, the number of samples required for that specific laboratory to achieve specified error rate objectives is determined. As a result, each laboratory produces unbiased compliance data. Also, compliance test error rates are identical for all laboratories conducting the test. Moreover, the data used to estimate laboratory bias and precision are similar to the QC measurements typically recommended for every analytical program. In summary, the intralaboratory approach appears, in general, to provide a greater degree of control over compliance test error rates while using QC resources more efficiently than the approach requiring interlaboratory QC data.

REFERENCES

Satterthwaite, F.E. (1946), "An Approximate Distribution of Estimates of Variance Components", Biometrics Bulletin, Vol. 2, pp. 110-114.

Youden, W.J.; and Steiner, E.H. (1975), Statistical Manual of AOAC. Association of Official Analytical Chemists, Washington, D.C.

DISCUSSION

George T. Flatman
U.S. Environmental Protection Agency

Dr. Bertram Price has something worth saying and has said it well in his paper entitled, "Quality Control Issues in Testing Compliance with a Regulatory Standard: Controlling Statistical Decision Error Rates."

The Environmental Protection Agency is emphasizing "Data Quality Objectives." Dr. Price has expressed the most important of these objectives in his title, "Controlling Statistical Decision Error Rates." The paper is timely for EPA because it demonstrates how difficult the statistics and the implementation are for data quality objectives.

In Section 1...Introduction, an "interlaboratory study approach" is suggested for establishing "bias and variability criteria." This is theoretically valid but may not be workable in practice. In contract laboratory programs, standards are in a much cleaner matrix (distilled water instead of leachate) and sometimes run on cleaner instruments that have not just run dirty specimens. Standards or blank samples cannot avoid special treatment by being blind samples since they are in a different matrix than the field samples. Thus, in practice, the same matrix and analytical instruments must be used to make "interlaboratories study" an unbiased estimate of the needed "bias and variability criteria." Both the theory and the implementation must be vigorously derived.

In Section 2...Statistical Models the enumeration of the components of variation is important for both theory and practice. More precise enumeration of variance components than the mutually exclusive and jointly exhaustive theory of "between and within" is needed for adequate sampling design. I agree with Dr. Price that "simplification, limiting the number of variance components, does not limit the generality of subsequent results," but I suggest it makes biased or aliased data collection more probable. For example, the Superfund Interlaboratories Studies of the Contract Labs has identified the calibration variance of the analytical instrument as the largest single component of longitudinal laboratory (or interlaboratories) variance. If this component of variation is not enumerated explicitly, I suggest this component of variance could be omitted, included once, or included twice. If all the field samples and lab replicate analyses were run between recalibrations of the analytical instrument, the recalibration variance would be omitted from the variances of the data. If the analytical instrument were recalibrated in the stream of field samples and between lab replicate analyses, the recalibration variance would be aliased with both the sample and lab variances, and thus added twice into the total variance. With these possible analyses scenarios the recalibration component of variance could be either omitted or included twice. This potential for error can be minimized through the vigorous modeling of all the process sources of variation in the components of variance model. This is not a criticism of the paper but it is a problem for the implementation of this paper by EPA's data quality objectives.

Section 3...Statistical Test is very important because it specifically states the null and alternative hypotheses with their probability alpha of type I error and probability beta of type II error. This may appear pedantic to the harried practitioner, but due to the importance of the decision is absolutely essential to data quality objectives. Dr. Price's alternative hypothesis and his beta-algebra is complicated by EPA's interpretation of the law, "no exceedence of background values or concentration limits" (40 CFR part 264). This requires an interval alternative hypothesis

$$H_1: \mu > \mu_0$$

rather than Dr. Price's point hypothesis

$$H_1: \mu = \mu_0 + D.$$

Lawyers should be more aware of how they increase the statistician's work. Beta is a function or curve over all positive D.

I think it is important to mention in any environmental testing that beta is more critical or important than in historical hypotheses testing. Classically the hypotheses are formulated so that a type II error is to continue with the status quo when in fact a new fertilizer, brand of seed potato, etc., would be better. Thus, the loss associated with the type II error is low and its probability of occurrence can be large (e.g., 20 percent) in agricultural experiments. This is not true in environmental hypotheses testing. The hypotheses usually make a type II error the misclassification of "dirty" as "clean" with a loss in public health and environmental protection. Thus, beta representing the probability of this loss in public health and environmental protection should be set arbitrarily low like alpha (1% or 5%).

Sections 4 and 5...Sample Size Requirements derive equations for numbers of field samples and lab replicates as a function of cost and variances. The formulas digitize the process for precise decisions between number of field samples and number of lab replicates. The formulas indicate that an analysis instrument like GCMS, because of its high incremental analysis cost and low variance requires few replications (K=1), but other analysis instruments such as radiation counters may not. These formulas have a practical value because of the diversity of analysis instruments and pollutants.

Section 5...Sample Size Requirements: Values of Variance Components Unknown detail the rigors of variance components estimation through unknown degrees of freedom and non-central t-distribution.

It might be asked, is not only the sum of variances needed for testing or "quality assurance" (i.e., rejection of outliers). This is true, but "quality improvement" requires the estimation of each component of variance. The analysis is more meaningful and usable if the individual components have an estimate.

Section 6...Discussion and Conclusions state that interlaboratories QC model (variable effects) and intralaboratory QC model (fixed effects) "lead to compliance tests that satisfy specified decision error rate objectives." This theoretical position of the paper is confirmed by the empirical findings of the Superfund Interlaboratories Comparison of the Contract Laboratories. This study found that within-lab variance is of corresponding magnitude to between-lab variance. The appropriate test and model should be used that correspond to the use of one lab or more than one lab in the actual chemical analysis of the data.

In conclusion, Dr. Bertram Price has rigorously presented the algorithms and the problems for "Controlling Statistical Decision Error Rates." This paper enumerates the statistical problems in applying hypothesis testing to real world data. Unfortunately, hypotheses testing is made deceptively simple in many textbooks and the true complexity is discovered in practice through the expensive consequences of a wrong decision. The serious problems discussed in Dr. Price's paper are needed to sober the superficial use of "alphas, betas, and other probabilities" in data quality objective statements. The paper is a timely and vigorous summary of components of variance modeling and hypotheses testing.

Acknowledgments: The discussant wishes to thank Forest Garner and Evangelos Yfantis for their advice, review, and insight gained from Superfund interlaboratories testing.

Notice: Although the thoughts expressed in this discussion have been supported by the United States Environmental Protection Agency, they have not been subject to Agency review and therefore do not necessarily reflect the views of the Agency and no official endorsement should be inferred.

ON THE DESIGN OF A SAMPLING PLAN TO VERIFY COMPLIANCE WITH EPA STANDARDS FOR RADIUM-226 IN SOIL AT URANIUM MILL-TAILINGS REMEDIAL-ACTION SITES

R.O. Gilbert, Pacific Northwest Laboratory; M.L. Miller, Roy F. Weston,
Inc.; H.R. Meyer, Chem-Nuclear Systems, Inc.

1.0 INTRODUCTION

The United States government is required under the Uranium Mill Tailings Radiation Control Act (U.S. Congress Public Law 95-604, 1978) to perform remedial actions on inactive uranium mill-tailings sites that had been federally supported and on properties that had been contaminated by the tailings. The current Environmental Protection Agency (EPA) standard for ^{226}Ra (henceforth denoted by Ra) in soil (EPA, 1983) requires that remedial action must be taken if the average concentration of Ra in surface (0- to 15-cm) soil over any area of 100 square meters exceeds the background level by more than 5 pCi/g, or if the average exceeds 15 pCi/g for subsequent 15-cm thick layers of soil more than 15 cm below the surface. Since there are many thousands of 100 square-meter areas that must be evaluated, the soil sampling plan should be as economical as possible while still meeting the intent of the regulations.

After remedial action at a site has been conducted, the field sampling procedure that has been used to determine whether the EPA standard was met was to first grid the entire site into 10-m by 10-m plots. Then, in each plot, 20 plugs of surface soil were collected and physically mixed together from which a single 500-g composite sample was withdrawn and assayed for Ra. If this measurement was $\geq$ 5 pCi/g above background, then additional remedial action was required. Recently, based on cost considerations and the study described in Section 2.0, the number of soil plugs per composite sample was reduced from 20 to 9.

In this paper we discuss a verification acceptance-sampling plan that is being developed to reduce costs by reducing the number of composite soil samples that must be analyzed for Ra. In Section 2.0 we report on statistical analyses of Ra measurements on soil samples collected in the windblown mill-tailings flood plain at Shiprock, NM. These analyses provide guidance on the number and size of composite soil samples and on the choice of a statistical decision rule (test) for the acceptance-sampling plan discussed in Section 4.0. In Section 3.0, we discuss the RTRAK system, which is a 4-wheel-drive tractor equipped with four Sodium-Iodide (NaI) gamma-ray detectors. The RTRAK is being developed for measuring radionuclides that indicate the amount of Ra in surface soil. Preliminary results on the calibration of these detectors are presented.

2.0 PERCENT ACCURACY OF MEANS AND PROBABILITIES OF DECISION ERRORS

In this section we statistically analyze Ra measurements of composite soil samples collected from the windblown mill-tailings flood-plain region at Shiprock, NM. This is done to evaluate the impact on probabilities of false positive and false negative decision errors resulting from reducing the number of soil plugs per composite soil sample from 21 to 9 or 5 and from collecting 1, 2, or 3 composite samples per plot. We also consider how these changes affect the accuracy of estimated mean Ra concentrations.

2.1 FIELD SAMPLING DESIGN

The Shiprock study involved collecting multiple composite soil samples of different sizes from 10 plots in the flood-plain region after an initial remedial action had occurred. Five sizes of composite samples were collected; those formed by pooling either 5, 8, 9, 16, or 21 plugs of soil.

Figure 1 shows the windblown mill-tailings flood-plain region and the location of ten 30-m by 30-m study areas from which composite soil samples were collected. Eight- and 16-plug composite samples were formed by pooling soil plugs that were collected over the ten 30-m by 30-m areas according to the three sampling patterns shown in the lower half of Fig. 2. The 5-, 9-, and 21-plug composite samples were formed by pooling soil plugs collected from only the central 10-m by 10-m plot in each 30-m by 30-m area using the three patterns shown in the upper half of Fig. 2.

Up to nine composite samples of each type were formed in each of the ten areas. Each composite sample of a given type used the same pattern that had been shifted slightly in location. For example, referring to Fig. 2, the 21-plug composite sample number 1 in a given 10-m by 10-m plot was formed by pooling soil plugs collected at the 21 positions numbered 1 in the plot. This design allowed replicate composite samples of a given type to be collected without altering the basic pattern that would be used in practice.

Each soil plug was collected to a depth of 15 cm using a garden trowel. The plugs collected for a given composite sample were placed in a bucket and mixed vigorously by stirring and shaking. The composite sample analyzed for Ra consisted of about 500 g of the mixed soil.

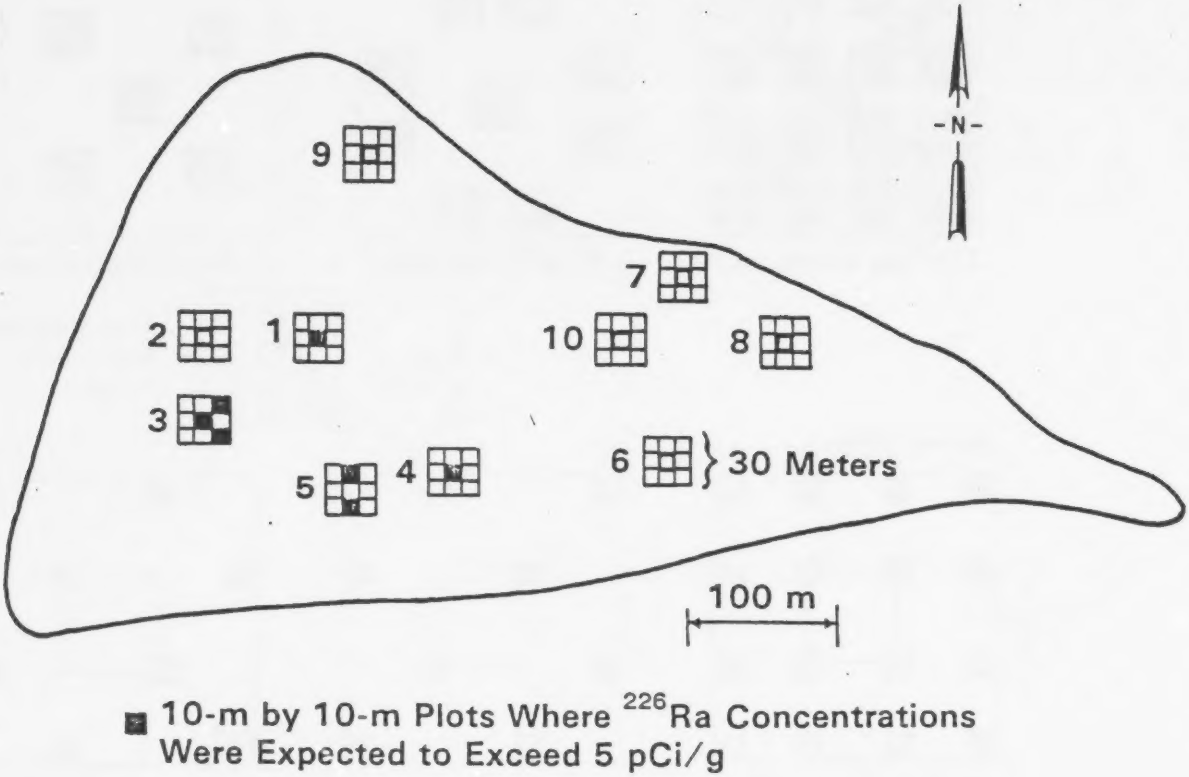

-N-
9
7
2
1
10
8
3
6
30 Meters
5
4
100 m
■ 10-m by 10-m Plots Where ^{226}Ra Concentrations Were Expected to Exceed 5 pCi/g

FIGURE 1. Location of the Ten 30-m by 30-m Areas in the Windblown Mill-tailings Flood Plain Region at Shiprock, New Mexico, Within Which Multiple-composite Soil Samples were Collected Following Initial Removal of Surface Soil.

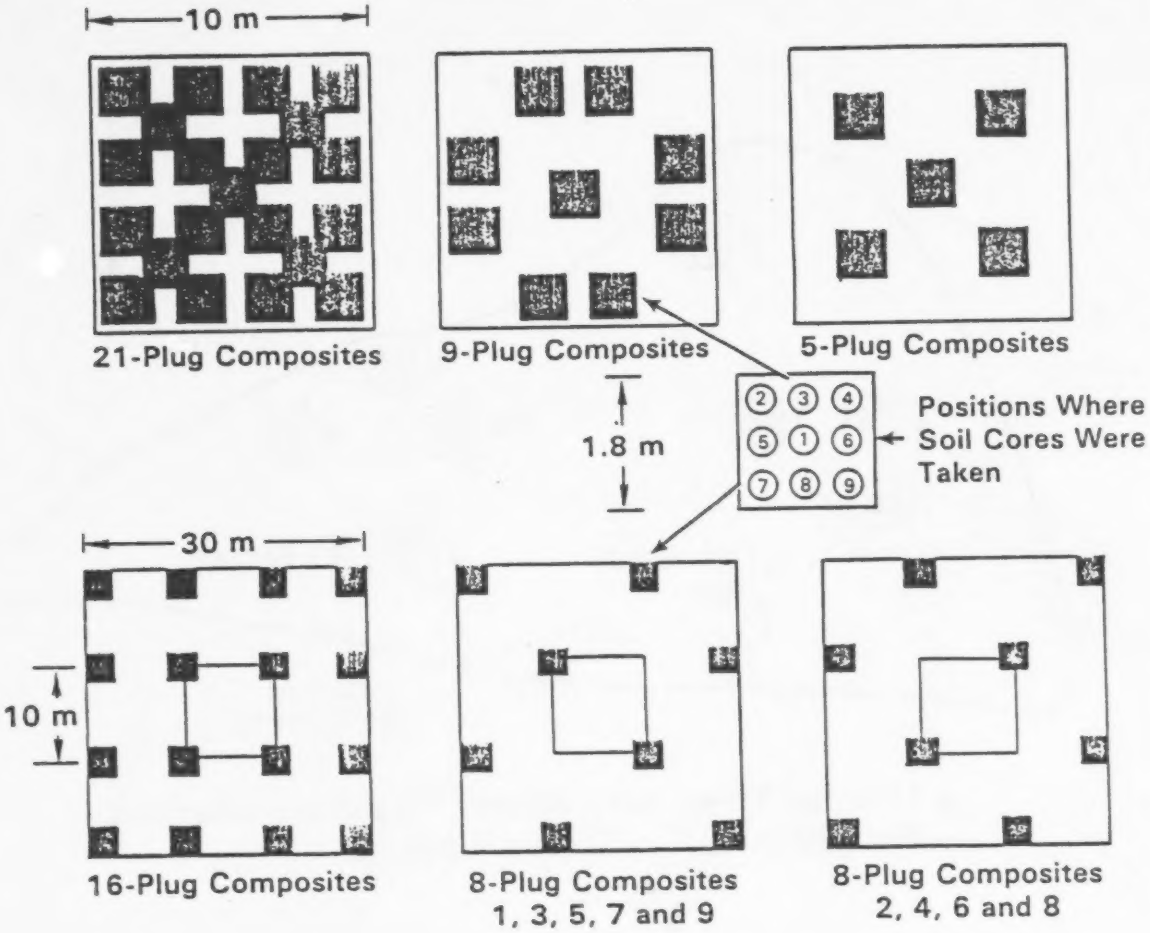

10 m
21-Plug Composites
9-Plug Composites
5-Plug Composites
1.8 m
② ③ ④
⑤ ① ⑥
⑦ ⑧ ⑨
Positions Where Soil Cores Were Taken
30 m
10 m
16-Plug Composites
8-Plug Composites 1, 3, 5, 7 and 9
8-Plug Composites 2, 4, 6 and 8

FIGURE 2. Sampling Patterns for 5-, 8-, 9-, 16-, and 21-plug Composite Soil Samples Collected From Ten 30-m by 30-m Areas in the Windblown Mill-tailings Flood Plain at Shiprock, New Mexico.

2.2 DESCRIPTION OF THE DATA

The Ra measurements for the composite samples are plotted in Figs. 3, 4, and 5. The figures also give the arithmetic mean, $\bar{x}$, the standard deviation, s, and the number of replicate composite samples, n. We wish to determine the extent to which the true standard deviation, σ, increases when fewer than 21 plugs are used to form a composite sample. To avoid confusion, we point out that Figs. 4 and 5 indicate that Ra measurements of most 5-, 9-, and 21-plug samples from Areas 1, 3, and 4 are larger than measurements for the 8- and 16-plug samples from those areas. This is believed to have occurred because the soil in the central 10-m by 10-m plot (from which 5-, 9-, and 21-plug composite samples were formed) had higher concentrations of Ra than the soil in the 30-m by 30-m areas from which the 8- and 16-plug samples were formed (see Fig. 1).

Measurements for Areas 8, 9, and 10 were below 5 pCi/g (Fig. 3) and the standard deviations ranged from 0.2 to 0.8 pCi/g, with no apparent trends in s with increasing number of plugs per sample. The data in Fig. 4 indicates that 5-plug sample data sets may be more skewed than those for 9- or 21-plug samples, at least for some plots. The measurements for Areas 1, 4, and 7 (Fig. 5) had higher means and were more variable than those for the areas in Figs. 3 and 4. In Fig. 6 are plotted the values of s from Figs. 3, 4, and 5 to show more clearly the changes in s that occurred as the number of plugs per composite sample changed.

2.3 ESTIMATING AND MODELING CHANGES IN STANDARD DEVIATIONS

In this section we first estimate the changes in σ that occur as the number of plugs per composite sample decreases from 21 to a smaller number. Then a model for these changes is developed for use in later sections.

A simple model for the ratio of standard deviations is obtained by assuming that measurements of Ra in individual soil plugs are uncorrelated, than the soil plugs are thoroughly mixed together before the 500-g aliquot is removed, and that the standard deviation between soil plugs does not change as the

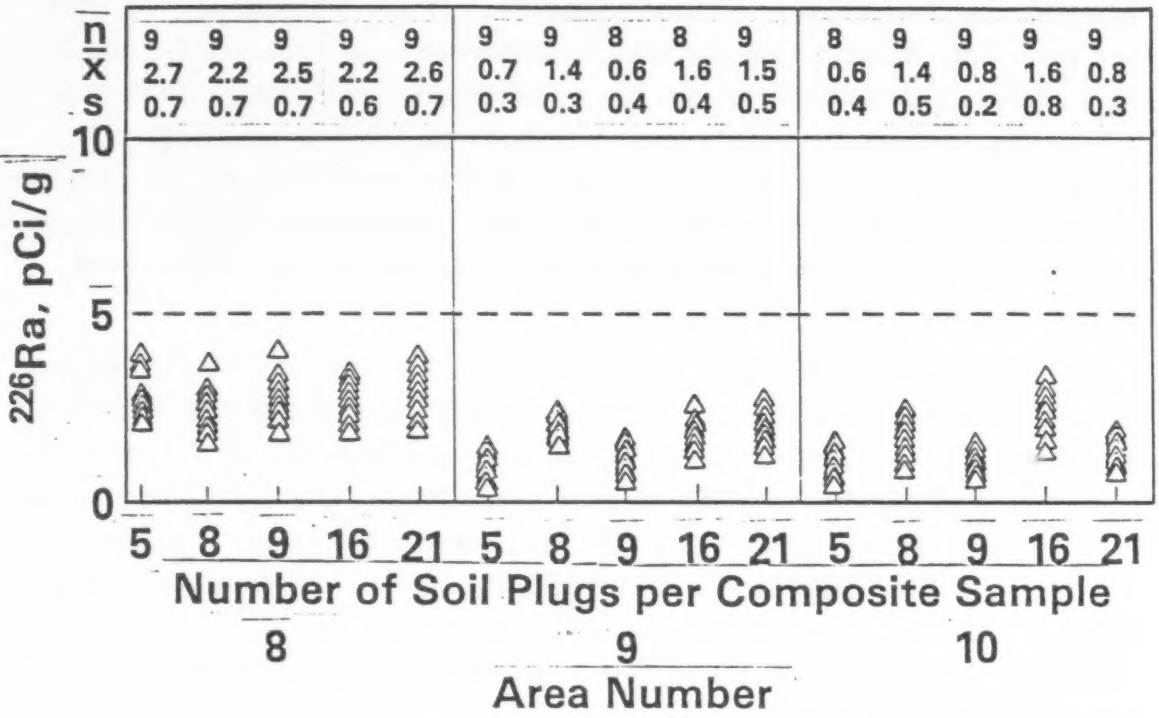


FIGURE 3. ^{226}Ra Measurements (pCi/g) of 5-, 8-, 9-, 16-, and 21-plug Composite Soil Samples Taken from Areas 8, 9, and 10 in the Windblown Mill-tailings Flood Plain at Shiprock, New Mexico. $\bar{x}$ and s are the Arithmetic Mean and Standard Deviation of the n Measurements for each Data Set.

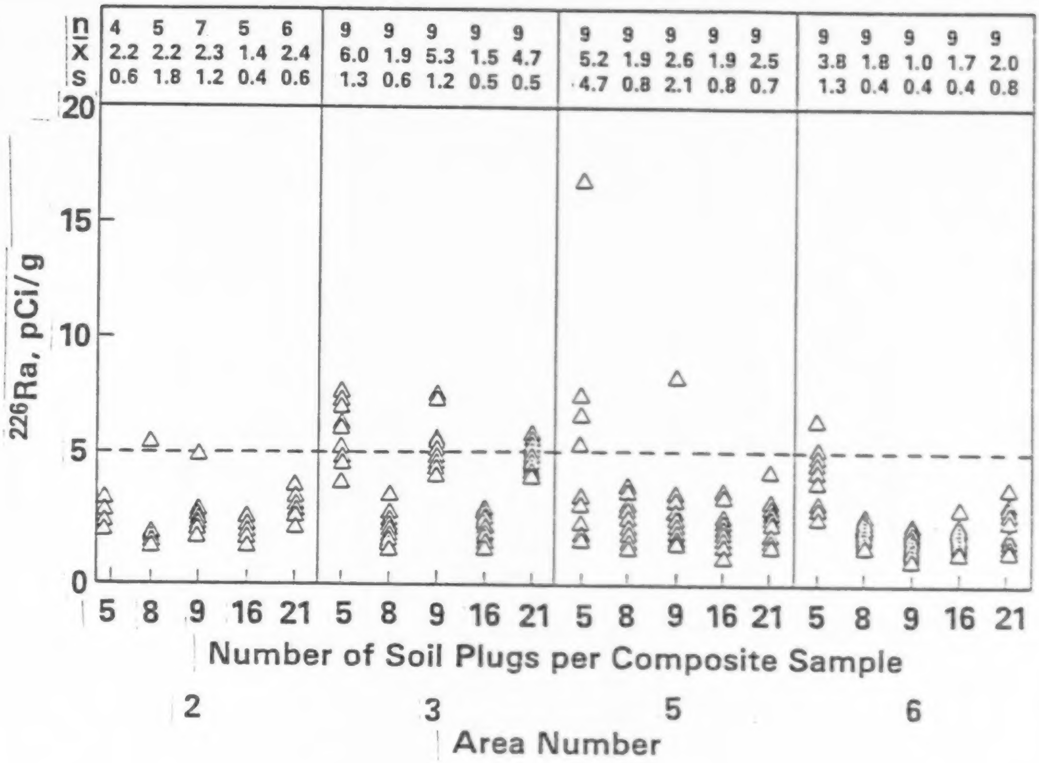


FIGURE 4. ^{226}Ra Measurements (pCi/g) of 5-, 8-, 9-, 16-, and 21-plug Composite Soil Samples Taken from Areas 2, 3, 5, and 6 in the Windblown Mill-tailings Flood Plain at Shiprock, New Mexico. $\bar{x}$ and s are the Arithmetic Mean and Standard Deviation of the n Measurements for each Data Set.

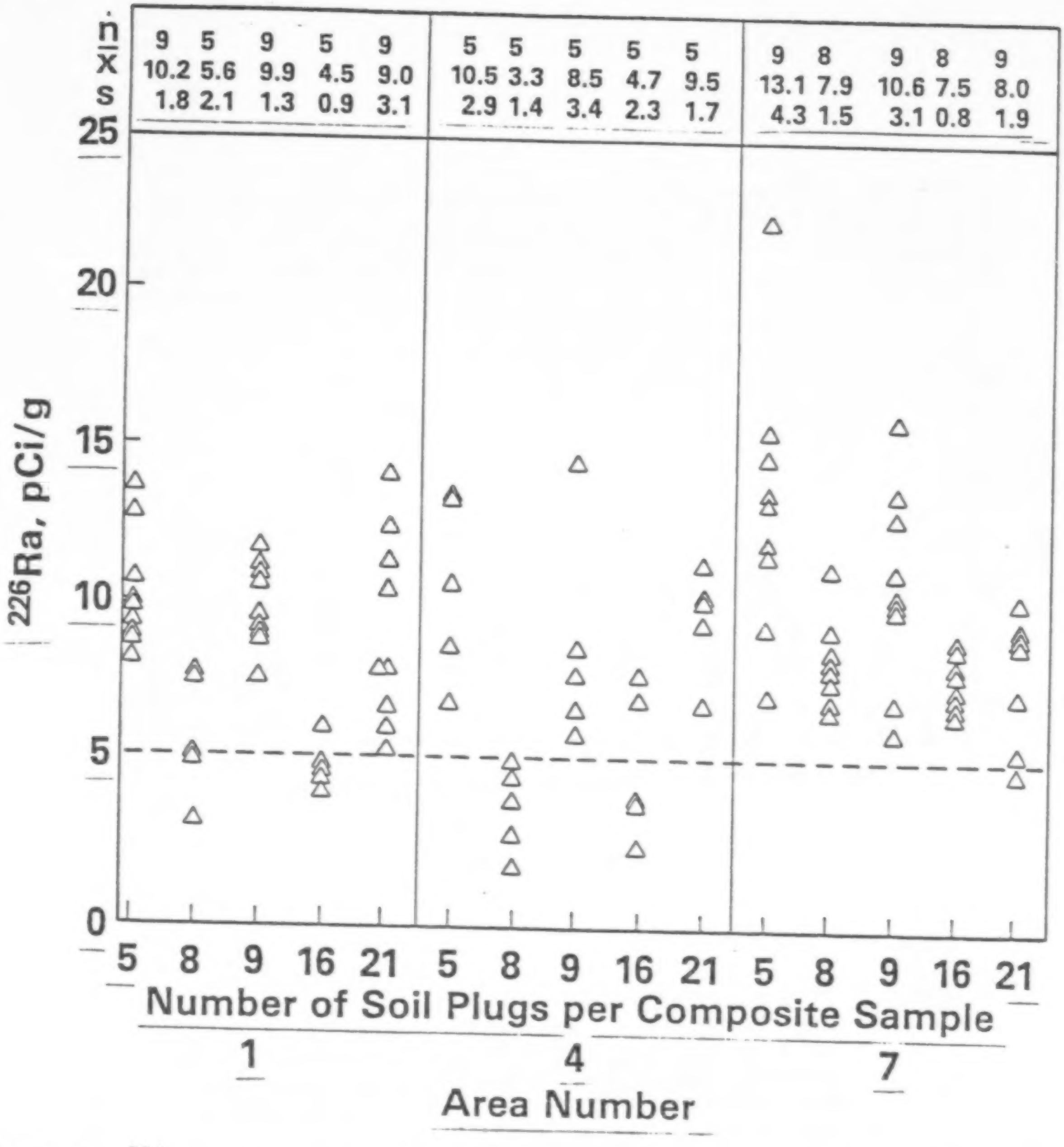


FIGURE 5. ^{226}Ra Measurements (pCi/g) of 5-, 8-, 9-, 16-, and 21-plug Composite Soil Samples Taken from Areas 1, 4, and 7 in the Windblown Mill-tailings Flood Plain at Shiprock, New Mexico. $\bar{x}$ and s are the Arithmetic Mean and Standard Deviation of the n Measurements for each Data Set.

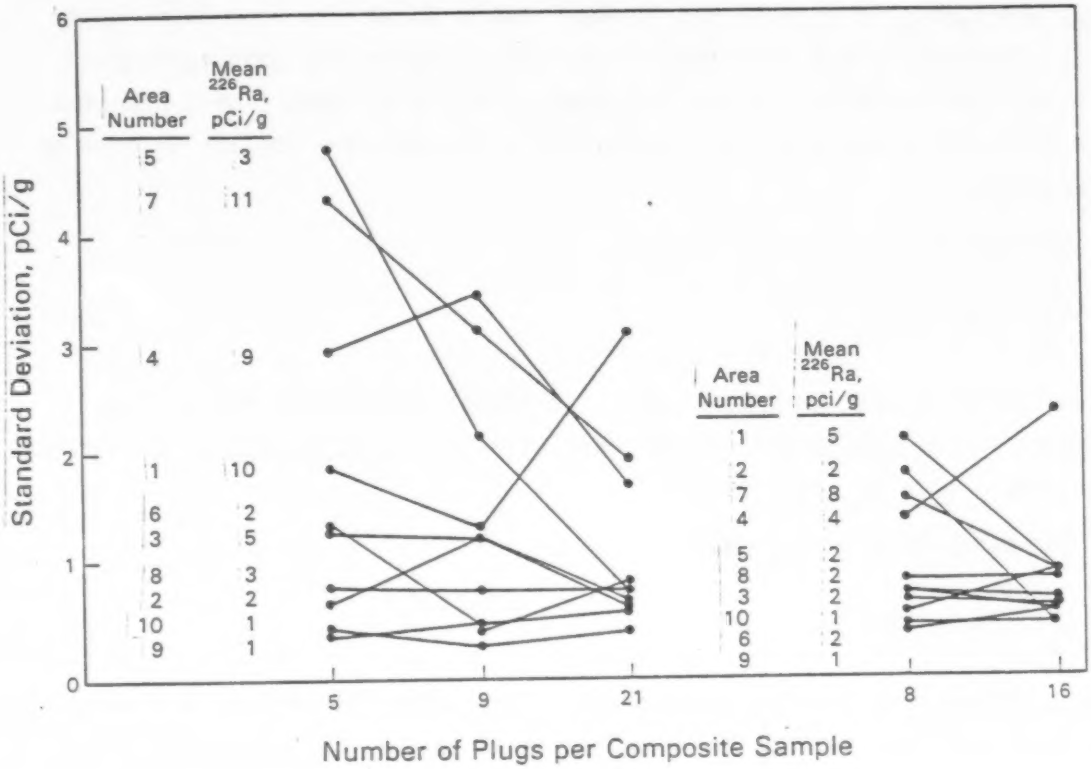


FIGURE 6. Standard Deviations of Multiple Composite Samples from Areas 1 Through 10 at the Windblown Mill-tailings Flood Plain at Shiprock, New Mexico. Mean ^{226}Ra Concentrations for each Area are Given to Illustrate that Areas with Lower Average Concentrations tend to have Smaller and More Stable Standard Deviations.

sampling pattern (see Fig. 2) changes. Under these assumptions we have the model

$$\sigma_{p_1}/\sigma_{p_2} = (\sigma'/\sqrt{p_1})/(\sigma'/\sqrt{p_2}) = (p_2/p_1)^{1/2} \qquad (1)$$

where σ' is the standard deviation for individual soil plugs.

Table 1 (column 6) gives values of Eq. 1 for comparison with estimated geometric means (GMs) and arithmetic means (AMs) of the ratios s_9/s_{21}, s_5/s_{21}, s_5/s_9, and s_8/s_{16} (columns 2 and 4) where the s values are from Figs. 3, 4, and 5. The modeled and estimated values are in reasonably good agreement. (Note that the estimated ratios in columns 2 and 4 of Table 1 were computed after excluding Areas 9 and 10 since those areas had very low and uniform Ra measurements.)

Solving Eq. (1) for σ_{p_1} gives

$$\sigma_{p_1} = \sigma_{p_2}(p_2/p_1)^{1/2} \qquad (2)$$

This equation is used here to predict the standard deviation for p_1-plug composite samples using the standard deviation for p_2-plug composite samples (σ_{p_2}), where $p_2 = 21$ and $p_1 < 21$.

The model used for σ_{p_2} was

$$\sigma_{p_1} = 0.10 + 0.23\mu \qquad (3)$$

where μ is the true mean Ra concentration (including background) for the plot. This model was used because the standard deviation of 21-plug samples tends to increase as the mean Ra concentration increases. This can be seen in Fig. 7 where we have plotted, for each of the 10 areas at Shiprock, the value of s versus the mean Ra measurement for composite samples formed from 5, 9, and 21 plugs of soil. Least-squares linear regression lines were fit to the three sets of data. The least-squares line for the 21-plug samples is the basis for the model in Eq. (3).

TABLE 1. Comparing Estimated and Predicted Ratios of Standard Deviations for Composite Samples Formed From Different Numbers of Soil Plugs.

Ratio of Standard Deviations	Estimated Ratios[+] Computed Using Data from Areas 1 through 8				Predicted[**] Ratios Computed Using Equation 1
	Geometric Mean (GM)	Geometric Standard Error[++] (GSE)	Arithmetic Mean (AM)	Standard Error (SE)	
σ_9/σ_{21}[*]	1.3	1.3	1.6	0.3	1.53
σ_5/σ_{21}	1.7	1.3	2.2	0.7	2.05
σ_5/σ_9	1.3	1.2	1.5	0.3	1.34
σ_8/σ_{16}	1.4	1.3	1.7	0.5	1.41

---

[*] σ_j = true standard deviation of j-plug composite samples.

[**] Computed as $(p_2/p_1)^{1/2}$, where p_1 and p_2 are the smaller and larger number of soil plugs per composite sample, respectively.

[+] Areas 9 and 10 were excluded because of their very low and uniform ^{226}Ra measurements.

[++] GSE = exp $(s_\ell/\sqrt{\sigma_n})$ where s_ℓ is the estimated standard deviation of the natural logarithms (n = 8).

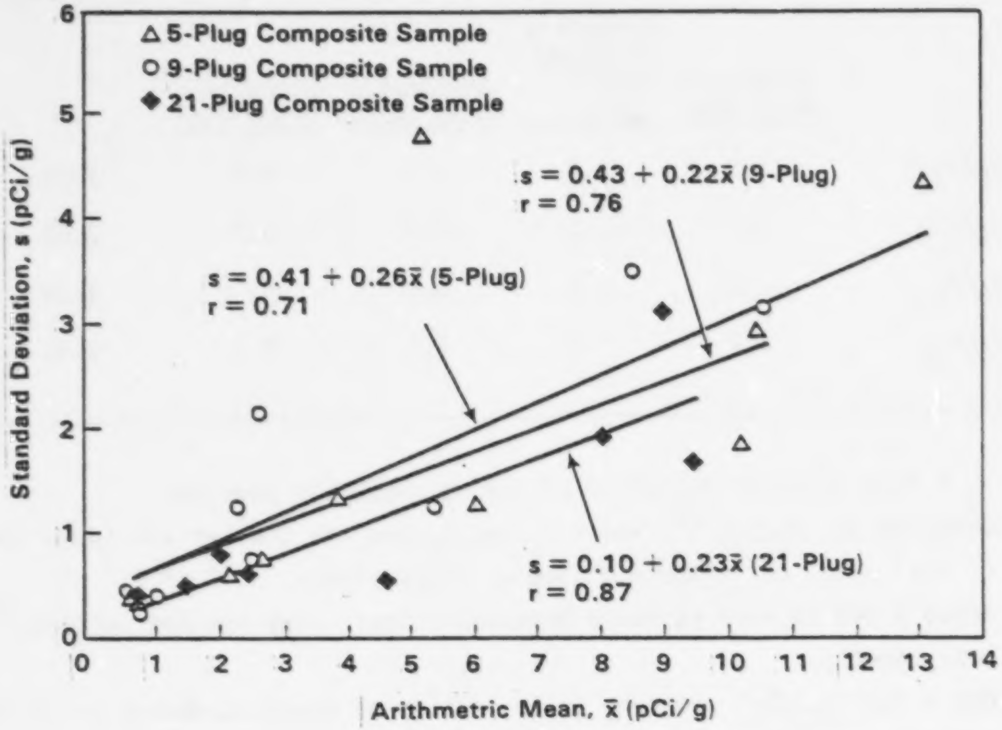


FIGURE 7. Least-Squares Linear Regression Lines Relating the Standard Deviation of Replicate Composite Samples from a Plot to the Estimated Mean Concentration of ^{226}Ra for the Plot.

Substituting Eq. (3) in Eq. (2) gives

$$\sigma_{p_1} = (0.10 + 0.23\mu)(p_2/p_1)^{1/2} \qquad (4)$$

which is the model used here to predict the standard deviation of p_1-plug composite samples, where $p_1 < 21$. The equations for 5- and 9-plug samples in Fig. 7 were not used to predict standard deviations because of the relatively small correlations (r) obtained for those data.

2.4 PERCENT ACCURACY OF ESTIMATED MEAN Ra CONCENTRATIONS

Using Eq. (4) and assuming that Ra measurements of composite samples are normally distributed, the following formula was used to estimate the percent accuracy with which the post-remedial-action mean Ra concentration for a plot at Shiprock would be estimated with specified confidence:

$$\text{Percent Accuracy} = 100\, Z\, (0.10+0.23\mu)(p_2/p_1)^{1/2}/(\mu\sqrt{n}), \qquad (5)$$

where Z equals 1.96 or 1.28 if 95% or 80% confidence, respectively, is required, n is the number of p_1-plug composite samples collected in the plot and averaged together to estimate the plot mean, and μ is the true plot mean. Eq. (5) is based on the usual formula for estimating the number of samples required to estimate a mean with prespecified relative accuracy and confidence; see, e.g., Gilbert (1987, p. 33).

In Fig. 8 are plotted values of Eq. (5) for 80% and 95% confidence, p_1 = 5, 9, and 21 plugs, n = 1 and 2 composite samples per plot, and for μ ranging from 1 to 10 pCi/g. To illustrate the meaning of Fig. 8, consider the plotted value for 95% confidence, p_1= 9, n = 2, and μ = 8. If two 9-plug samples are from a 10-m by 10-m plot that has a true mean concentration of 8 pCi/g (including background), then we can be 95% sure that the arithmetic mean of the two measurements will fall within about 51% of the true mean.

The curves in Fig. 8 show that approximately doubling the number of plugs per sample increases the percent accuracy by 20 to 25 percentage points. Also, the increase in percent accuracy is negligible if more than 4 composite samples are used.

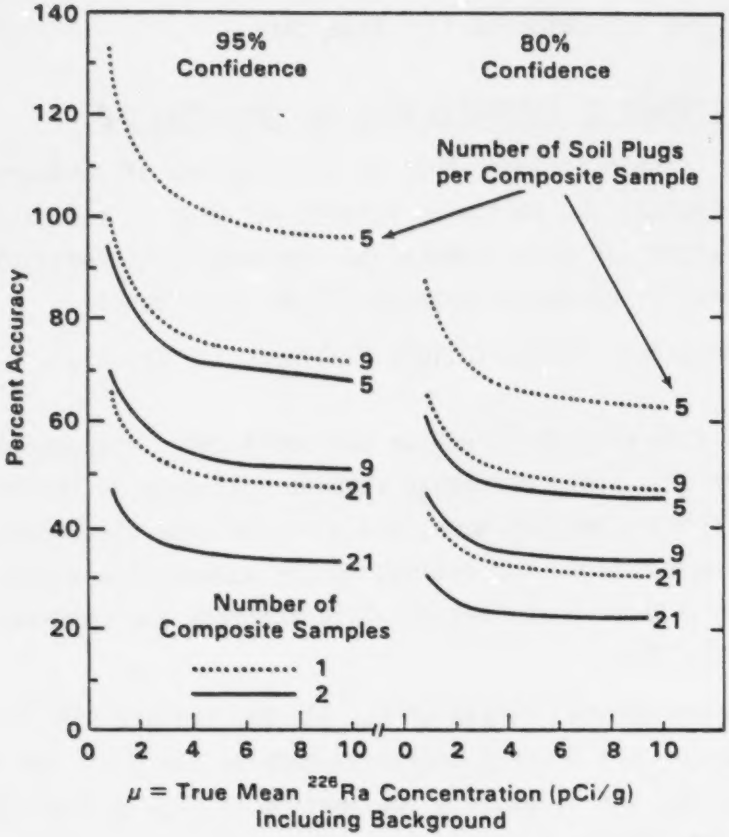


FIGURE 8. Percent Accuracies for Estimated Mean ^{226}Ra Concentrations in Surface Soil for 10-m by 10-m Plots at the Shiprock, New Mexico Site.

By dividing Eq. 5 when $p_2 = 21$ and $p_1 < 21$ by Eq. 5 when $p_2 = p_1 = 21$ we obtain $(21/p_1)^{1/2}$, which is the factor by which the percent accuracy of 21-plug composite samples is multiplied to get the percent accuracy of p_1-plug samples. This formula gives 1.5 and 2.0 when $p_1 = 9$ and 5, respectively. Notice that this factor is not study-site dependent since it does not depend on μ or σ.

2.5 PROBABILITIES OF REMEDIAL ACTION DECISION ERRORS

In this section the increase in remedial-action decision errors as the number of plugs per sample declines is quantified. These results are obtained assuming: (1) that Eq. 4 is an appropriate model for the variance of p_1-plug composite samples ($p_1 < 21$), (2) the estimated Ra mean concentration for a plot based on p_1-plug composite samples withdrawn from the plot is normally distributed, and (3) the mean Ra background concentration is known.

The probabilities of making remedial action decision errors are computed for three different decision rules:

Decision Rule 1

Take additional remedial action if $\bar{x}' + 1.645\ \sigma_{p_1}/\sqrt{n}$ (the upper 95% confidence limit on the true plot mean) exceeds 5 pCi/g above background, where $\bar{x}'$ is the estimated mean concentration (above background) for the plot based on n p_1 - plug composite samples.

Decision Rule 2

Take additional remedial action if $\bar{x}'$ exceeds 5 pCi/g above background.

Decision Rule 3

Take additional remedial action if $\bar{x}' - 1.645\ \sigma_{p_1}/\sqrt{n}$ (the lower 95% confidence limit on the true plot mean) exceeds 5 pCi/g above background.

Among these three rules, Rule 1 offers the greatest protection to the public because the probabilities of taking additional remedial action are greater than for rules 2 or 3. Rule 3 will result in fewer decisions to take remedial action than rules 1 or 2 for plots with true mean Ra concentrations near 5 pCi/g above background. Hence, Rule 3 will tend to reduce costs of remedial action. Rule 2 is a compromise strategy in that the probabilities of taking remedial action fall between those for Rules 1 and 3.

Let us define β to be the probability that a statistical test will indicate additional remedial action is needed. When Decision Rule 1 is used, the probability β is obtained by computing:

$$Z_\beta = \frac{(5-\mu')(np_1/21)^{1/2}}{\sigma_{21}} - 1.645, \tag{6}$$

where 5 is the EPA limit, μ' is the true plot mean above background, σ_{21} is the standard deviation of 21-plug composite samples given by Eq. (3), p_1 is the number of soil plugs used to form each of the n composite samples from distribution. Z_β is then referred to tables of the cumulative normal distribution to determine β.

For Decision Rule 2, the same procedure is used except that Eq. (6) is computed with the constant 1.645 replaced by zero. For Decision Rule 3, the negative sign before 1.645 in Eq. (6) is replaced by a positive sign.

We computed β for various values of μ' when the background Ra concentration was assumed to be 1 pCi/g (the approximate background value for the windblown flood plain at the Shiprock site) when n = 1, 2, or 3, and p_1 = 5, 9, or 21. The results when n = 1 are plotted in Fig. 9, and the results for one, two, or three 9-plug composite samples are plotted in Fig. 10.

These figures indicate that:

1. Decreasing the number of plugs per composite sample increases the probability of incorrectly deciding additional remedial action is needed.

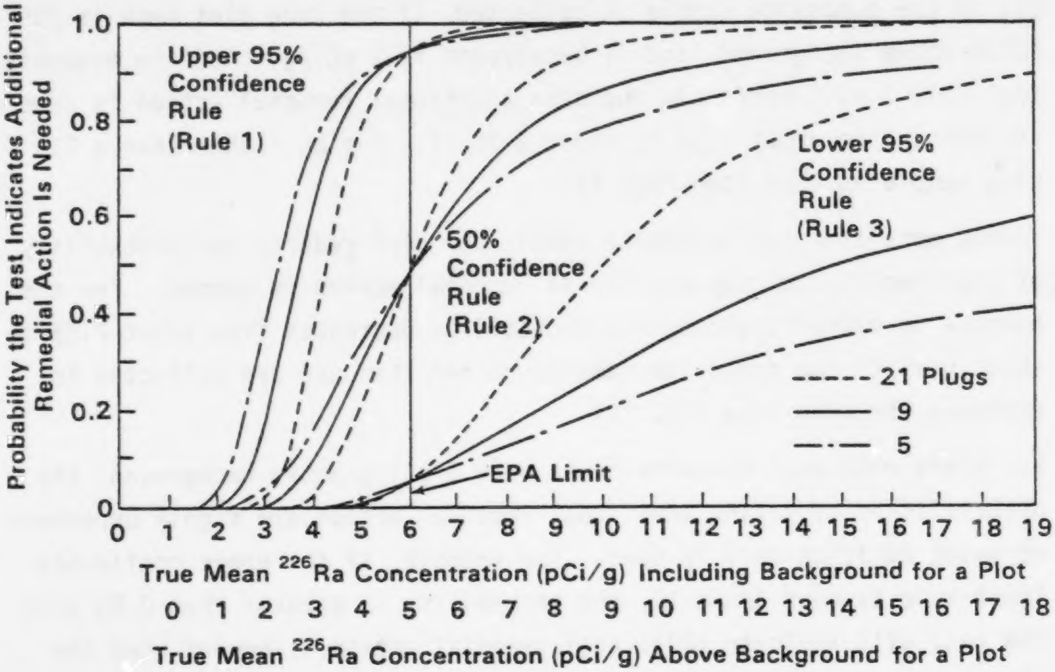


FIGURE 9. Probabilities of Taking Additional Remedial Action in a Plot for Three Decision Rules When One 500-g Sample from a Composite Sample Composed of Either 21, 9, or 5 Soil Plugs from the Plot is Measured for ^{226}Ra.

For example, if the upper confidence limit rule is used (Rule 1), if one composite sample is collected, if the true mean for the plot is 3 pCi/g above background, and if background is 1 pCi/g, then the probability the rule will indicate additional remedial action is needed increases from about 0.40 to about 0.65 if a 9-plug rather than a 21-plug composite sample is used to estimate the plot mean (see Fig. 9).

2. Decreasing the number of plugs per composite sample increases the probability of incorrectly deciding additional remedial action is not needed. For example, if the lower confidence limit rule is used (Rule 3), if one composite sample is collected, if the true plot mean is 10 pCi/g above background, and if background is 1 pCi/g, then the probability that Rule 3 will correctly indicate additional remedial action is needed decreases from about 0.60 to about 0.30 if a 9-plug rather than a 21-plug sample is used (see Fig. 9).
3. Taking more than one composite sample per plot reduces the probability of incorrectly deciding additional remedial action is needed. For the example in number 1 above, the probability decreases from about 0.65 to about 0.45 if two composite samples rather than one are collected to estimate the mean (see Fig. 10).
4. For plots with mean concentrations near 5 pCi/g above background, the probabilities of taking additional remedial action are highly dependent on which decision rule is used. For example, if the upper confidence limit rule is used (Rule 1), the probability is greater than 0.95 that the test will indicate additional remedial action is needed when the plot has a mean Ra concentration greater than 5 pCi/g above background. But if the lower confidence limit rule (Rule 3) is used, and one 21-plug composite sample is collected, the probability that the test will indicate additional remedial action is needed does not reach 0.95 until the true plot mean is about 20 pCi/g above background. Rule 2 falls between these two extremes. It achieves a 0.95 probability (for one or more 21-plug samples) when the true mean above background is about 9 or 10 pCi/g (see Fig. 9).

The three decision rules may find application at different times in the remedial action process. The upper confidence limit rule seems most appropriate

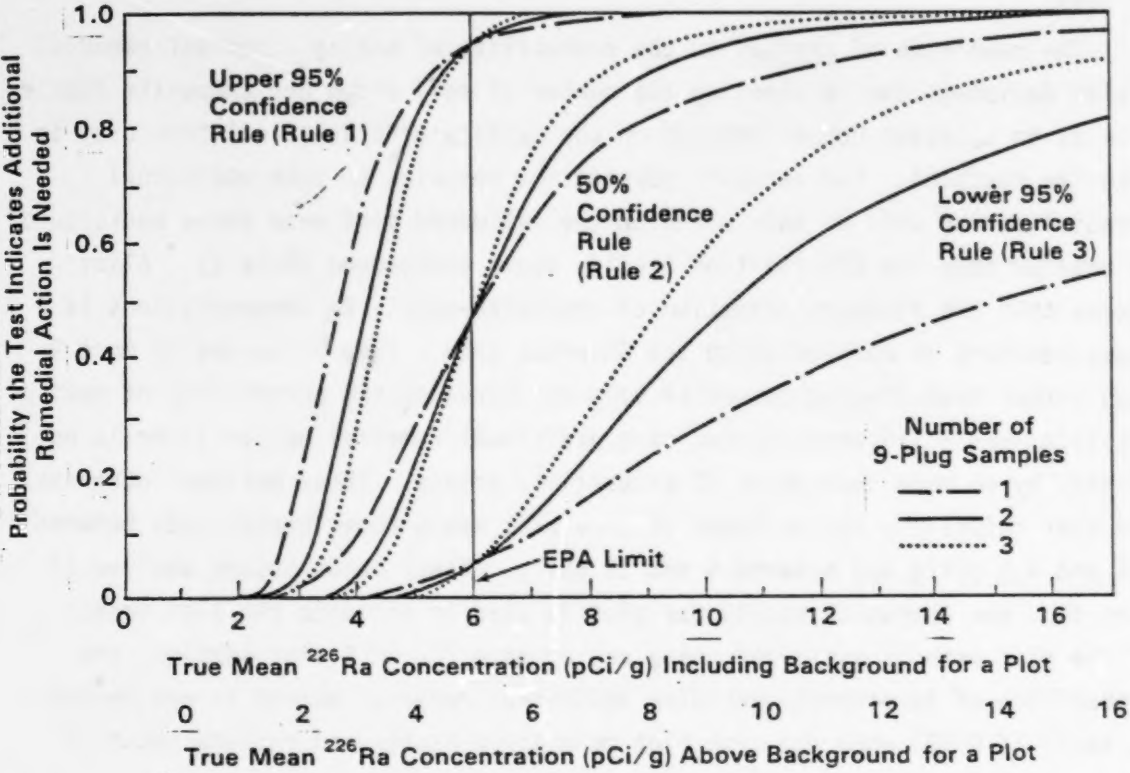


FIGURE 10. Probabilities of Taking Additional Remedial Action in a Plot for Three Decision Rules if One, Two, or Three 500-g Samples from a Composite Sample Composed of 9 Soil Plugs are Measured for ^{226}Ra.

at initial stages when it may be prudent to assume that the plot is contaminated until proven otherwise. The "price" of using this rule is increased remedial action costs for plots that have true mean concentrations just under 5 pCi/g above background. The lower confidence limit rule is more appropriate for plots that are strongly believed to have already been cleaned to below the EPA limit. Using this rule, the probability of taking additional remedial action is less than 0.05 when the true plot mean is 5 pCi/g above background or less.

The magnitude of changes in the probability of making incorrect remedial action decisions due to changing the number of soil plugs per composite sample from 21 to a lesser number depends on the particular statistical test used to make the decision. For example, suppose the decision to take additional remedial action will be made whenever the estimated plot mean above background is greater than the EPA limit of 5 pCi/g above background (Rule 2). Also, assume that the standard deviation of composite-sample Ra concentrations is a known constant as modeled using the Shiprock data. Then using one or more 9-plug rather than 21-plug composite samples increases the probability of making decision errors (incorrectly deciding additional remedial action is or is not needed) by no more than about 17 probability points. These maximum increases are over relatively narrow bands of true plot means above background; between 2.5 and 4.5 pCi/g and between 6 and 13 pCi/g. These bands become smaller if more than one composite sample per plot is used to estimate the plot mean. If the plot mean is estimated using one or more 21- or 9-plug samples, the probability of incorrectly deciding additional remedial action is not needed is small (≤ 0.05) when the true plot mean above background exceeds about 15 pCi/g.

If Rules 1 and 3 are to yield and probabilities shown in Figs. 9 and 10 the true standard deviation for the plot must be given by Eq. (4). At contaminated sites where this model does not apply, special soil sampling studies could be conducted to determine whether Eq. (4) or some other model is applicable. Alternatively, if several composite samples are collected from each plot then the standard deviation could be estimated directly for each plot using those data. Then upper or lower confidence limits would be computed using the t distribution rather than the normal distribution [see

Exner et al. (1985) for an application of the upper confidence limit test]. Use of the t distribution will generally give more decision errors, which is the price paid when the standard deviation must be estimated. If the mean background Ra concentration is estimated, this will also increase the standard deviation and hence the probabilities of making decision errors.

As concerns the comparison of 21-, 9-, and 5-plug samples, the increase in probabilities of decision errors as the number of plugs per composite sample is reduced is, on the whole, about the same as shown in Figs. 9 and 10 when the standard deviation, σ_{p_1}, was assumed known. This conclusion is based on probabilities of decision errors we obtained using the noncentral t distribution and the methods in Wine (1964), pp. 254-260). These results are shown in Fig. 11 for the case of two composite samples per plot.

2.6 EXPECTED NUMBER OF DECISION ERRORS

The expected number of plots at a remediated site that are misclassified as needing or not needing additional remedial action depends on the probabilities of making decision errors and on the frequency distribution of the true plot means. Fig. 12 shows the frequency distribution of estimated Ra means for 1053 plots at the Shiprock floodplain site that had undergone an initial remedial action (removal of soil). Each mean was estimated by the measurement of one 20-plug composite sample from the plot. Fig. 12 shows that 83 plots had estimated means that exceeded the EPA standard of 1 pCi/g above background (6 pCi/g).

We assume for illustration purposes that the histogram in Fig. 12 is the distribution of true plot means. (When the RTRAK system becomes operational, it is expected that, following remedial action, all plots will have Ra concentrations below the EPA limit. Hence, the distribution in Fig. 12 may be a worst case distribution.) Under this assumption we wish to determine the effect of using 9 rather than 21 plugs of soil per composite sample on the expected number of plots that are misclassified. Let n_i be the number of plots in the ith frequency class, Q be the number of classes, and p_i be the probability of a decision error for a plot with true mean in the ith class using a chosen decision rule. Then $E = \sum n_i p_i$ is the expected number of misclassified plots for the decision rule.

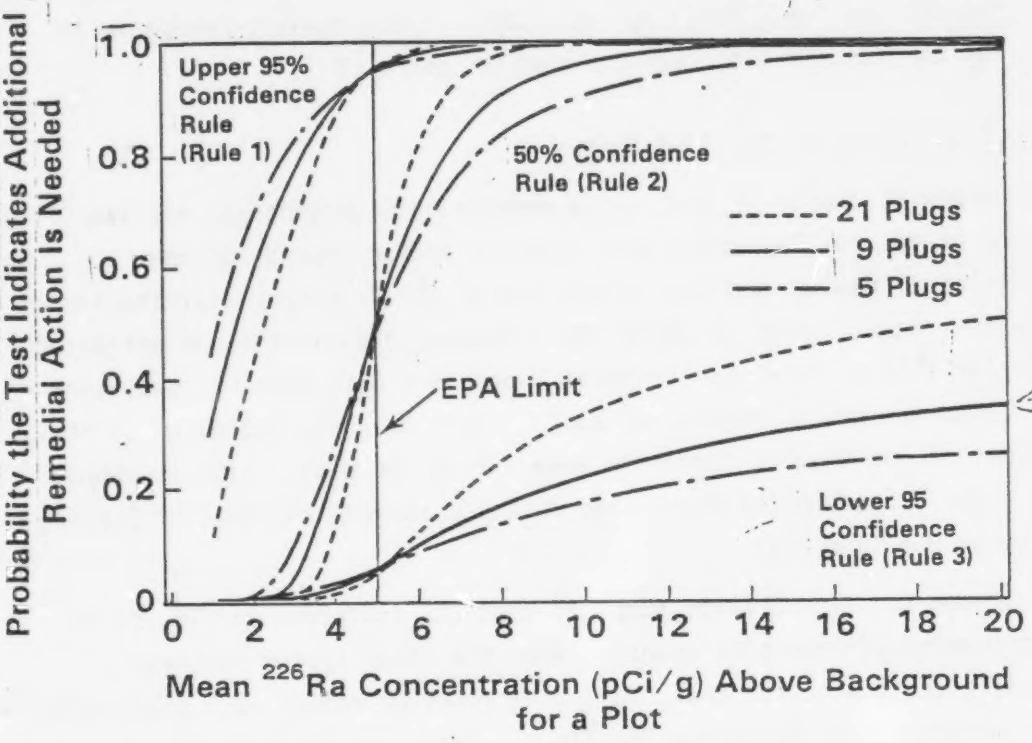


FIGURE 11. Probabilities of Taking Additional Remedial Action in a Plot for Three Decision Rules if Two 21-, 9-, or 5-Plug Composite Soil Samples are Collected and the t Test is Used to Make Decisions.

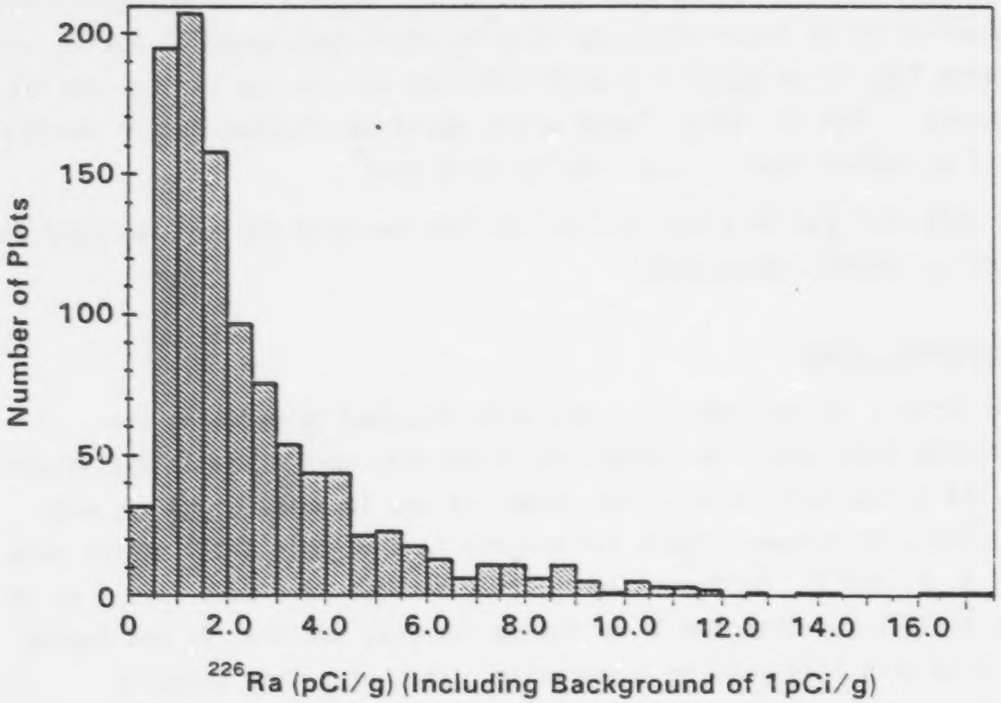


FIGURE 12. Frequency Distribution of Estimated Mean ^{226}Ra Concentrations (pCi/g) in Surface Soil following Initial Remedial Action for 1053 10-m by 10-m Plots in the Windblown Mill-tailings Flood Plain at Shiprock, New Mexico.

First, we computed E for the 970 plots in the Q = 12 classes in Fig. 12 that had means less than 6 pCi/g, i.e., for plots that met the EPA standard. Using the probabilities in Fig. 9 for Rule 2 of incorrectly deciding to take additional remedial action, we found that E = 27.4 and 40.2 for 21- and 9-plug samples, respectively. Hence, the use of a single 9-plug rather than a single 21-plug composite in each plot would result in an expected 13 more plots undergoing unneeded additional remedial action.

Next, we computed E for the 83 plots in Fig. 12 that had means greater than 6 pCi/g, i.e., for plots needing additional cleanup. Using Rule 2 and the probabilities of incorrectly deciding no additional remedial action was needed from Fig. 9, we found E = 12.95 and 19.5 for 21- and 9-plug samples, respectively. That is, about 7 more plots would not receive needed remedial action if 9- rather than 21-plug samples were used.

We note that the 83 plots in Fig. 12 that exceeded the EPA standard were subsequently further remediated:

2.7. LOGNORMAL MODEL

The results in Sections 2.3 - 2.6 were obtained by modeling the untransformed data under the assumption those data were normally distributed. We used the W statistic to test for normality and lognormality (see, e.g. Gilbert (1987) or Conover (1980) for descriptions of this test) of the data in Figs. 5, 6, and 7. We found that 21-plug samples were more likely to be normally distributed than the 9- or 5-plug samples, and that 9- and 5-plug samples were more likely to be lognormally distributed than normally distributed. Also, the increase in the standard deviation as the mean increases (see Fig. 7) indicates that the lognormal distribution may be a better model for these data than the normal distribution.

In this section we investigate the extent to which the probability results in Section 2.5 would change if the lognormal distribution rather than the normal distribution was appropriate. To do this, the natural logarithms of the data in Figs. 3, 4, and 5, were computed and a model was developed for the standard deviation of the logarithms. We found that after deleting the data for plots 9 and 10 (the standard deviation of the logarithms (s_y) for these plots were about twice as large as for the remaining eight plots) there was no statistically significant linear relationship between s_y and the mean of the logarithms. This indicates that the lognormal distribution may be a reasonable model, at least for plots with concentrations at the level of those in plots 1 through 8. The pooled standard deviation of the logarithms for plots 1-8 was 0.4, 0.37, and 0.3 for 5-, 9-, and 21-plug samples, respectively.

The probabilities of taking additional remedial action were computed for Rule 2 for the case of one, two, or three 5-, 9-, and 21-plug samples using these modeled standard deviations. This was done by computing

$$Z_\beta = (\ln 5 - \ln \mu')/\sigma_y$$

and referring Z_β to the standard normal distribution tables, where σ_y equalled 0.4, 0.37, and 0.3 for 5-, 9-, and 21-plug samples, respectively.

We found that for 9-plug samples, the false-positive error probabilities for the lognormal case differed by less than two probability points from those for the normal case for all mean Ra concentrations less than the EPA limit. Differences in the false-negative rates were as large as 8 probability points for mean concentrations between 8 and 10 pCi/g above background for the case of one 9-plug composite sample per plot. These results, while limited in scope, suggest that the false-positive and false-negative error probabilities in Section 2.5 may be somewhat too large if the lognormal distribution is indeed a better model for the Ra data than the normal distribution.

3.0 RTRAK AND ITS CALIBRATION

The RTRAK is a 4-wheel-drive tractor equipped with four Sodium-Iodide (NaI) detectors, their supporting electronics, an industrial-grade IBM PC, and a commercial microwave auto-location system. The detectors are independently mounted on the front of the tractor and can be hydraulically lifted and angled. Bogey wheels support the detectors to maintain a distance of 12 inches from the ground during monitoring. Each detector has a tapered lead shield that restricts its field of view to about 12 inches, with overlap between adjacent detectors. The RTRAK will take gamma-ray readings while moving at a constant speed of 1 mph. When a reading above a prespecified level is encountered, red paint is sprayed on the ground to mark these "hot spots". The automatic microwave locator system provides x-y coordinates with the count data. This will permit real-time map generation to assist in control of contamination excavation. Preliminary data indicate that the RTRAK should be able to detect Ra in soil at concentrations less than 5 pCi/g. Further tests of the RTRAK's detection capabilities are underway.

The proper calibration of the RTRAK detectors is important to the success of the remedial-action effort. The Na(I) detectors detect selected radon daughter gamma peaks that are related to Ra. Hence, the RTRAK detectors do not directly measure Ra, the radionuclide to which the EPA standard applies. Radon is a gas, and the rate that it escapes from the soil depends on several factors including soil moisture, source depth distribution, soil randon emanating fraction, barometric pressure, soil density, and soil composition. The calibration of the detectors must take these variables into account so that randon daughter gamma peaks can be accurately related to Ra concentrations under field conditions.

A field calibration experiment near the Ambrosia Lake, NM, mill-tailings pile was recently conducted as part of the effort to develop a calibration procedure. In this experiment the RTRAK accumulated counts of ^{214}Bi (Bismuth) for approximately 2-second intervals while traveling at 1 mph. Red paint was sprayed to mark the locations and distances traveled for each time interval. For each detector, from 3 to 5 surface soil samples were collected down the centerline of each scanned area (Fig. 13). Then, for each of these areas,

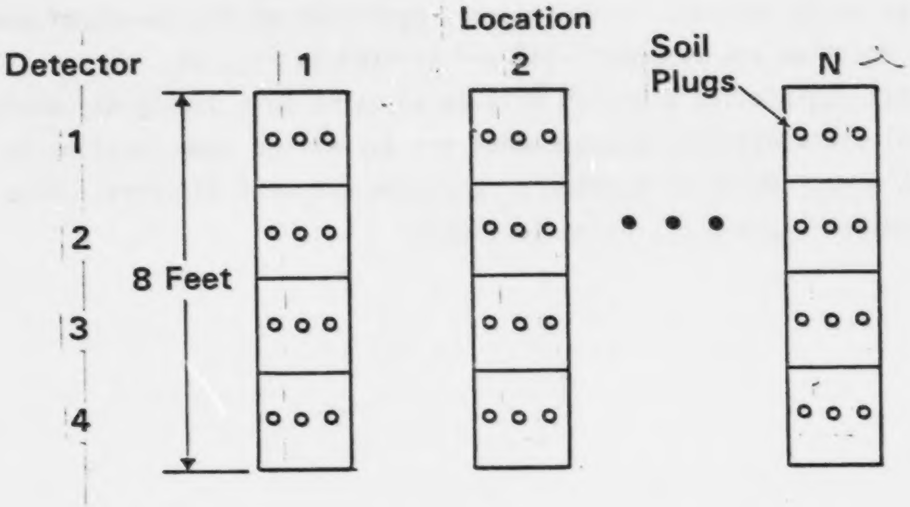

Location
Detector
1
2
N
Soil Plugs
1
2
3
4
8 Feet

FIGURE 13. Pattern of Soil-Sample Locations and RTRAK Detector Readings for Obtaining Data to Calibrate the Detectors.

these samples were mixed and a ~ 500-g aliquot was removed and sealed in a metal can that was assayed for Ra within a few days and then again following a 30-day waiting period to permit equilibrium to be established between Rn and ^{214}Bi.

The data and the fitted least-squares linear regression line are displayed in Fig. 14. The data for the 4 detectors have been combined into one data set because there were no important differences in the 4 separate regression lines. Also shown in Fig. 14 are the 90% confidence intervals for predicted Ra individual measurements. The regression line and limits in Fig. 14 were obtained by first using ordinary least-squares regression on the ln-transformed data. Then the equation was exponentiated and plotted in Fig. 14. It is expected that this calibration equation will be adjusted on a day-by-day basis by taking several RTRAK-detector measurements per day at the same location in conjunction with measurements of barometric pressure and soil moisture. This adjustment procedure is presently being developed.

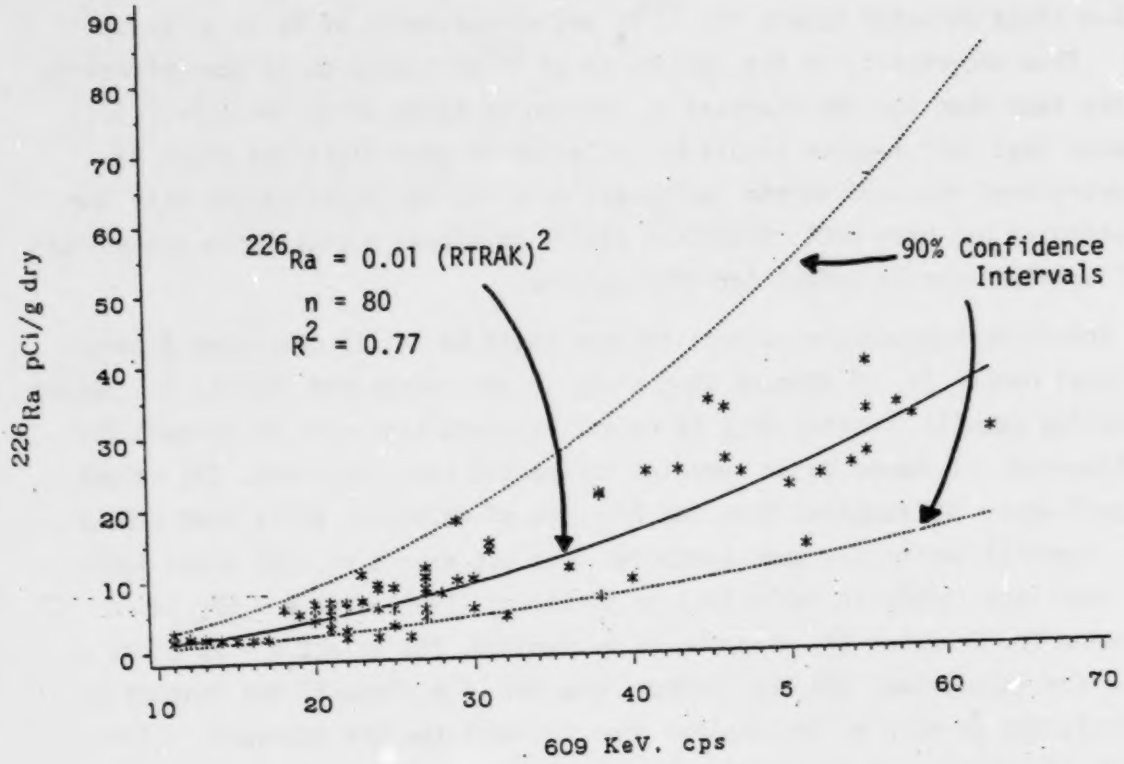


FIGURE 14. Least-squares Regression Line for Estimating ^{226}Ra Concentrations (pCi/g) in Surface Soil Based on RTRAK-Detector Readings of Bi-214 (609 Kev).

4.0 COMPLIANCE ACCEPTANCE SAMPLING

As illustrated by Fig. 14, there is not a perfect one-to-one correspondence between RTRAK detector counts for ^{214}Bi and measurements of Ra in aliquots of soil. This uncertainty in the conversion of ^{214}Bi counts to Ra concentrations, and the fact that the EPA standard is written in terms of Ra concentrations, suggests that soil samples should be collected in some plots and their Ra concentrations measured in the laboratory as a further confirmation that the EPA standard has been met. Schilling (1978) developed a compliance acceptance-sampling plan that is useful for this purpose.

Schilling's procedure as applied here would be to (1) determine (count) the total number (N) of 10-m by 10-m plots in the remediated region, (2) select a limiting (small) fraction (P_L) of defective plots that will be allowed (if undiscovered) to remain after remedial action has been completed, (3) select the confidence (C) required that the fraction of defective plots that remain after remedial action has been conducted does not exceed P_L, (4) enter Table 1 in Schilling (1978) or Table 17-1 in Schilling (1982) with $D = NP_L$ to determine the fraction (f) of plots to be sampled, (5) select $n = fN$ plots at random for inspection, and (6) "reject" the lot of N plots if the inspection indicates one or more of the n plots does not meet the EPA standard. (The meaning of "reject" is discussed below.)

In Step 6, each of the n plots would be "inspected" by collecting three or four 9- or 21-plug composite soil samples and using these to conduct a statistical test to decide if the plot meets the EPA standard. The choice of three or four 9- or 21-plug samples is suggested by the results of our statistical analyses in Section 2.0 in the windblown mill-tailings flood plain region at Shiprock, NM.

Steps 4 and 5 can be simplified by using curves (Hawkes, 1979) that give n at a glance for specified N, P_L, and C. Also, the Operating Characteristic (OC) curves for this procedure (curves that give the probability of rejecting the lot [of N plots] as a function of the true fraction of plots that exceeds the standard) can be easily obtained using Table 2 in Schilling (1978) or Table 17-2 in Schilling (1982).

To illustrate the 6-step procedure above, suppose C = 0.90 and P_L = 0.05 are chosen, and that the remediated region contains N = 1000 plots. Then we find from Fig. 1 in Hawkes (1979) that n = 46 plots should be inspected. If all 46 inspected plots are found to be non-defective, we can be 100C = 90% confident that the true fraction of defective plots in the population of N = 1000 plots is less than 0.05, the specified value of P_L. If one or more of the n plots fail the inspection, then our confidence is less than 0.90.

As another example, suppose there are N = 50 plots in the remediated region of interest. Then, when C = 0.90 and P_L = 0.05, we find that n = 30 plots should be inspected. Small lots that correspond perhaps to subregions of the entire remediated region may be needed if soil excavation in these regions was difficult or more subject to error because of hilly terrain or other reasons.

The action that is taken in response to "rejecting the lot" may include collecting three or four 9- or 21-plug composite soil samples in adjacent plots surrounding the inspected plots that exceeded the EPA standard. The same statistical test as used previously in the original n plots would then be conducted in each of these plots. If any of these plots were contaminated above the EPA limit, they would undergo remedial action and gamma scans using the RTRAK system, and additional adjacent plots would be sampled, and so forth. The calibration and operation of the RTRAK NaI detectors would also need to be double checked to be sure the detectors and entire RTRAK system is operating correctly.

An assumption underlying Schilling's procedure is that no decision error is made when inspecting any of the n plots. However, inspection errors will sometimes occur since "inspection", as discussed above, consists of conducting a statistical test for each plot using only a small sample of soil from the plot. When inspection errors can occur, the fraction of defective plots is artificially increased, which increases the probability of rejecting the lot. To see this, let P denote the actual fraction of plots whose mean exceeds the EPA limit, let P_1 denote the probability of a false-positive decision on any plot (deciding incorrectly that additional remedial action is needed), and let P_2 denote the probability of a false-negative decision (deciding incorrectly that no additional remedial action is needed). Then, the effective fraction defective is $P_e = P_1(1-P) + P(1-P_2)$. For example, if $P_1 = P_2 = P = 0.05$, then $P_e = 0.05(0.95) + 0.05(0.95) = 0.095$ so that the compliance sampling plan will operate as if the true proportion of defective plots is 0.095 rather than 0.05. This means there will be a tendency to reject too many lots that actually meet the C and P_L specifications.

In Section 2.5 we saw, using Ra data from the Shiprock, NM, mill-tailings site, how P_1 and P_2 change with the statistical test used, the true mean concentration, the number of composite samples, and the amount of soil used to form each composite sample. If remedial action has been very thorough so that mean concentrations in all plots are substantially below the EPA limit, then the true fraction of defective plots, P, will be zero and $P_e = P_1$ (since $P = 0$) will be small. In that case, the probability of "rejecting the lot" using Shillings' compliance acceptance sampling plan will be small. As indicated above, this probability is given by the OC curve that may be obtained using Table 2 in Schilling (1978).

5.0 DISCUSSION

In this paper we have illustrated some statistical techniques for developing more cost-effective sampling plans for verifying that ^{226}Ra concentrations in surface soil meet EPA standards. Although the focus here was on ^{226}Ra in soil, these techniques can be used in other environmental cleanup situations. Because of the high cost of chemical analyses for hazardous chemicals, it is important to determine the number and type or size of environmental samples that will give a sufficiently high probability of making correct cleanup decisions at hazardous-waste sites. Also, it is clear from Section 2.5 above that when the level of contamination is close to the allowed maximum concentration limit, the probabilities of making correct cleanup decisions depend highly on the particular statistical test used to make decisions. Plots of probabilities such as given in Figs. 9, 10, and 11 provide information for evaluating which test is most appropriate for making remedial-action decisions.

A topic that is receiving much attention at the present time is the use of in-situ measurements to reduce the number of environmental samples that must be analyzed for radionuclides or hazardous chemicals. The RTRAK system discussed in this paper is an example of what can be achieved in the case of radionuclides in soil. Some in-situ measurement devices may only be sensitive enough to determine if and where a contamination problem exists. Other devices may be accurate enough to provide a quantitative assessment of contamination levels. In either case, but especially for the latter case, it is important to quantitatively assess the accuracy with which the in-situ method can measure the contaminant of interest. The regression line in Fig. 14 illustrates this concept.

It is hoped that this paper will provide additional stimulus for the use of statistical methods in the design of environmental sampling programs for the cleanup of sites contaminated with radionuclides and/or hazardous-waste.

6.0 REFERENCES

Conover, W. J. 1980. Practical Nonparametric Statistics, 2nd ed., Wiley, New York.

EPA 1983. Standard for Remedial Actions at Inactive Uranium Processing Sites; Final Rule (40 CFR Part 19 2). Federal Register 48 (3):590-604 (January 5, 1983).

Exner, J. H., W. D. Keffer, R. O. Gilbert, and R. R. Kinnison. 1985. "A Sampling Strategy for Remedial Action at Hazardous Waste Sites: Clean-up of Soil Contaminated by Tetrachlorodibenzo-p-Dioxin." Hazardous Waste and Hazardous Materials 2:503-521.

Hawkes, C. J. 1979. "Curves for Sample Size Determination in Lot Sensitive Sampling Plans", J. of Quality Technology 11(4):205-210.

Gilbert, R. O. 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, Inc., New York.

Schilling, E. G. 1978. "A Lot Sensitive Sampling Plan for Compliance Testing and Acceptance Inspection", J. of Quality Technology 10(2):47-51.

Schilling, E. G. 1982. Acceptance Sampling in Quality Control. Marcel Dekker, Inc., New York.

Wine, R. L. 1964. Statistics for Scientists and Engineers. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

DISCUSSION

Jean Chesson
Price Associates, Inc., 2100 M Street, NW, Washington, DC 20037

The presentation by Richard Gilbert provides a good illustration of several points that have been made by earlier speakers. My discussion is organized around three topics that have general applicability to compliance testing, namely, decision error rates, sampling plans, and initial screening tests.

Decision Error Rates

The EPA standard for Cleanup of Land and Buildings Contaminated with Residual Radioactive Materials from Inactive Uranium Processing Sites (48 FR 590) reads "Remedial actions shall be conducted so as to provide reasonable assurance that,....." and then goes on to define the requirements for concentrations of radium-226 in the soil. An objective way to "provide reasonable assurance" is to devise a procedure which maintains statistical Type II error rates at an acceptable level. A Type II error, or false negative, occurs when the site is declared in compliance when in fact it does not satisfy the standard. The probability of a Type II error must be low enough to satisfy EPA. On the other hand, the false positive (or Type I) error rate also needs to be kept reasonably low, otherwise resources will be wasted on unnecessary remedial action. The aim is to devise a compliance test that will keep Type I and II errors within acceptable bounds.

Developing a compliance test involves three steps. First, a plan for collecting data and a rule for interpreting it is specified. The paper considers several sampling plans and three decision rules for data interpretation. Second, the decision error rates are calculated based on a statistical model. In this case, the model involves a normal distribution, a linear relationship between the variance and mean for composite samples, and an assumption of independence between individual soil plugs making up the composite. The last two components of the model are based on empirical data. Third, the sensitivity of the estimated error rates to changes in the model assumptions should be investigated. This is particularly important if the same procedure is going to be applied at other sites. For example, if the estimated error rates are very sensitive to the model relating variance and mean, it will be necessary to verify the relationship at each site. Conversely, if the error rates are relatively insensitive to changes in the relationship, the compliance test could be applied with confidence to other sites without additional verification.

Sampling Plans

The sampling plan is an integral part of the compliance test. The paper illustrates how sampling occurs at several levels. There is the choice of plots within the site. The current plan involves sampling every plot. The proposed plan suggests sampling a subset of the plots according to an acceptance sampling plan. Then there is the choice of the number and type of samples. One or more samples may be collected per plot each composed of one or more soil plugs. Usually more than one combination will achieve the required decision error rates. The optimum choice is determined by the contribution of each type of sample to the total variance and by relative costs. For example, if variability between soil plugs is high but the cost of collecting them is low, and the measurement method is precise but expensive, it is advantageous to analyze composite samples composed of several soil plugs. If the measurement method is inexpensive, it may be preferable to analyze individual samples rather than composites.

Initial Screening Tests

The RTRAK is an interesting example of an initial screening test. Initial screening tests may be used by the regulated party to determine when the site is ready for the "real" compliance test, or they may be an integral part of the compliance test itself. In either case, the objective is to save costs by quickly identifying cases that are very likely to pass or to fail the clearance test. For example, if the RTRAK indicates that the EPA standard is not being met, additional remedial action can be taken before final soil sampling, thereby reducing the number of times soil samples are collected before the test is passed. If the initial screening test is incorporated in the compliance test, i.e., if a favorable result in the initial screening reduces or eliminates subsequent sampling requirements, then calculations of decision error rates must take this into account.

The "reasonable assurance" stated in the EPA rule is provided by an assessment of the decision error rates for the entire compliance test. The development and evaluation of a practical and effective multi-stage compliance test is a significant statistical challenge.

DISTRIBUTED COMPLIANCE: EPA AND THE LEAD BUBBLE

John W. Holley
Barry D. Nussbaum
U.S. EPA (EN-397F), 401 M St., S.W. Washington, D.C.

This paper discusses a particular class of strategies, "bubbles", for the management of human exposure to environmental hazards and examines an application of such strategies to the case of lead in gasoline. While gasoline is by no means the only source of environmental lead, for most of the population it has been the dominant source for many years and is certainly the most controllable source. Lead is not only toxic to people, it is also toxic to catalytic converters which are used on vehicles to reduce emissions of such conventional pollutants as carbon monoxide, hydrocarbons, and oxides of nitrogen. The twin objectives of protecting people from lead and from the conventional emissions of vehicles with lead-disabled catalysts led to the first Environmental Protection Agency (EPA) regulation of the substance in gasoline in 1979. This first regulation covered the total amount of lead allowed in each gallon of gasoline produced by a refinery when leaded and unleaded gasoline are considered together and averaged over a quarter. It also set up temporary standards at a less stringent level for small refiners. Without thinking of it in these terms, the Agency had taken the first steps toward recognizing the need for and implementing a "bubble" policy for lead. The paper will present some conceptual tools for discussing bubbles and then examine the application of this management approach to gasoline lead.

Bubbles--General Principles

In general, a bubble approach to environmental regulation may be thought of as any approach that aims at ensuring that environmental exposure to some pollutant is reduced or controlled "on the average" while accepting some variability across emitters in the magnitude of their contribution. "On the average" and "emitters" are ideas that obviously require further discussion.

Purposes of bubble regulations

Regulators may use bubbles for at least four reasons. First, they may allow institution of a stringent regulation that would be infeasible for each entity to meet, yet might be feasible for an industry as a whole. Second, bubbles make it possible to improve the flexibility of a regulation from the standpoint of the regulated entities and may thus lessen any negative economic impacts. The classic plant bubble is a case in point, providing for operating flexibility by regulating the pollution from the entire plant rather than that from each smokestack. Third, bubbles may improve the "fairness" of application of the burdens associated with a regulation. In this way regulators may mitigate the economic impact of an action upon firms that are somehow unusually sensitive to its provisions. The final reason for using a bubble approach is really derivative of the second and third. By minimizing and more fairly distributing the impact of a regulation, the drafter may make badly needed controls "possible" in a politico-economic sense. Thus the public health may be protected by a bubble regulation in a situation where the economic impact of a simpler regulation would make it politically impossible to achieve.

Logical elements of a bubble

A bubble regulation always has some dimension or set of dimensions along which compliance is distributed. The most obvious such dimension is space, and is illustrated again by reference to the plant/smokestack bubble. A lack of compliance in one location may be balanced off against greater than minimum compliance in another location. It is important in planning the implementation of a bubble regulation whether sources across which emissions are to be averaged are part of a single legally responsible entity (as in the plant model) or are each themselves separate corporate entities.

Time is another dimension along which compliance may be distributed. Almost all of our regulations are to some degree bubbles in this sense, since the dimension of time is always involved in our setting of compliance periods. Time even enters into our selection of the appropriate units (as in cubic feet per minute). This dimension becomes most important, though, in a situation where it is actively and intentionally manipulated in the design of the compliance strategy so as to achieve one or more of the objectives of bubbles that were mentioned above.

In addition to dimension, any successful bubble approach must have some thought given to what, for want of a better term, we may call an integrating medium. This medium must assure that the results of our allowing an uneven distribution of compliance across some dimension does not also result in sharp differences in the consequences of exposure across that same dimension. People in one area suffering from some kind of toxic exposure are afforded scant comfort by knowing that in consequence of their suffering the people in another area are not affected at all by the pollutant. So while we are attempting to achieve fairness in distributing the economic burdens of compliance among polluters, we must also consider the question of equity in exposure.

The integrating media in most bubbles are the classic air, water, soil, and food. Under some circumstances we may consider the human body to be an integrating medium, as in the case of pollutants whose effects are cumulative in the body over a lifetime. The air may mix the emissions from stack A and stack B so that the downwind victim experiences the average of the two. Certain pollutants may be diffused throughout a body of water in such a way that heavy emissions on one day may be balanced off

DISTRIBUTED COMPLIANCE: EPA AND THE LEAD BUBBLE

John W. Holley
Barry D. Nussbaum
U.S. EPA (EN-397F), 401 M St., S.W. Washington, D.C.

This paper discusses a particular class of strategies, "bubbles", for the management of human exposure to environmental hazards and examines an application of such strategies to the case of lead in gasoline. While gasoline is by no means the only source of environmental lead, for most of the population it has been the dominant source for many years and is certainly the most controllable source. Lead is not only toxic to people, it is also toxic to catalytic converters which are used on vehicles to reduce emissions of such conventional pollutants as carbon monoxide, hydrocarbons, and oxides of nitrogen. The twin objectives of protecting people from lead and from the conventional emissions of vehicles with lead-disabled catalysts led to the first Environmental Protection Agency (EPA) regulation of the substance in gasoline in 1979. This first regulation covered the total amount of lead allowed in each gallon of gasoline produced by a refinery when leaded and unleaded gasoline are considered together and averaged over a quarter. It also set up temporary standards at a less stringent level for small refiners. Without thinking of it in these terms, the Agency had taken the first steps toward recognizing the need for and implementing a "bubble" policy for lead. The paper will present some conceptual tools for discussing bubbles and then examine the application of this management approach to gasoline lead.

Bubbles--General Principles

In general, a bubble approach to environmental regulation may be thought of as any approach that aims at ensuring that environmental exposure to some pollutant is reduced or controlled "on the average" while accepting some variability across emitters in the magnitude of their contribution. "On the average" and "emitters" are ideas that obviously require further discussion.

Purposes of bubble regulations

Regulators may use bubbles for at least four reasons. First, they may allow institution of a stringent regulation that would be infeasible for each entity to meet, yet might be feasible for an industry as a whole. Second, bubbles make it possible to improve the flexibility of a regulation from the standpoint of the regulated entities and may thus lessen any negative economic impacts. The classic plant bubble is a case in point, providing for operating flexibility by regulating the pollution from the entire plant rather than that from each smokestack. Third, bubbles may improve the "fairness" of application of the burdens associated with a regulation. In this way regulators may mitigate the economic impact of an action upon firms that are somehow unusually sensitive to its provisions. The final reason for using a bubble approach is really derivative of the second and third. By minimizing and more fairly distributing the impact of a regulation, the drafter may make badly needed controls "possible" in a politico-economic sense. Thus the public health may be protected by a bubble regulation in a situation where the economic impact of a simpler regulation would make it politically impossible to achieve.

Logical elements of a bubble

A bubble regulation always has some dimension or set of dimensions along which compliance is distributed. The most obvious such dimension is space, and is illustrated again by reference to the plant/smokestack bubble. A lack of compliance in one location may be balanced off against greater than minimum compliance in another location. It is important in planning the implementation of a bubble regulation whether sources across which emissions are to be averaged are part of a single legally responsible entity (as in the plant model) or are each themselves separate corporate entities.

Time is another dimension along which compliance may be distributed. Almost all of our regulations are to some degree bubbles in this sense, since the dimension of time is always involved in our setting of compliance periods. Time even enters into our selection of the appropriate units (as in cubic feet per minute). This dimension becomes most important, though, in a situation where it is actively and intentionally manipulated in the design of the compliance strategy so as to achieve one or more of the objectives of bubbles that were mentioned above.

In addition to dimension, any successful bubble approach must have some thought given to what, for want of a better term, we may call an integrating medium. This medium must assure that the results of our allowing an uneven distribution of compliance across some dimension does not also result in sharp differences in the consequences of exposure across that same dimension. People in one area suffering from some kind of toxic exposure are afforded scant comfort by knowing that in consequence of their suffering the people in another area are not affected at all by the pollutant. So while we are attempting to achieve fairness in distributing the economic burdens of compliance among polluters, we must also consider the question of equity in exposure.

The integrating media in most bubbles are the classic air, water, soil, and food. Under some circumstances we may consider the human body to be an integrating medium, as in the case of pollutants whose effects are cumulative in the body over a lifetime. The air may mix the emissions from stack A and stack B so that the downwind victim experiences the average of the two. Certain pollutants may be diffused throughout a body of water in such a way that heavy emissions on one day may be balanced off against very light emissions on another day with the same effect as if daily emissions had been carefully held to an intermediate or average level.

Enforcement considerations

Measurement and/or sampling problems may arise with distributed compliance regulations that are rarely a problem with more conventional approaches. An example is a scheme for averaging automobile emissions across models or engine families that was considered by the Agency some years ago. Without a bubble approach the certification process is limited to determining whether each engine family meets a single standard. Under a bubble approach a whole set of issues arises around measuring the emission level of each family within some confidence limits--questions of sample size and design and distribution shape rear their heads. When these vehicles are tested to verify their in-use performance, statistical concerns again arise as we consider whether the manufacturer should be held responsible for the point estimate of certification emissions, the lower confidence limit (to provide maximum protection for the environment, or the upper confidence limit (to protect the manufacturer against unpleasant surprises that may be based upon sampling error). These statistical concerns clearly have sharply focussed policy and legal implications.

One effect of some distributed compliance schemes is to unintentionally compromise an environmental benefit which arises out of industry quality assurance provisions. In the simple situation where the manufacturer must meet a standard and face dire consequences for failing to do so, some "headroom" is likely to be left between the actual emission level and the somewhat higher standard. This gap benefits the environment to the extent of the manufacturer's intolerance of risk. A redesign of such an existing compliance scheme to a distributed compliance approach with payment of a monetary penalty for each ton of pollutant over the overall standard may lead to an increase in emissions by reducing the manufacturers' uncertainty, even though emissions overall remain under the statutory standard.

The enforcement of bubble regulations may cost more than would be the case for simpler alternatives. This is true because of the complexity of sampling and measurement and the administrative machinery needed to carry out enforcement. Where the bubble regulation provides significant benefits to the industry in the form of flexibility, but costs more to administer, the question arises as to whether the Agency or the industry should bear the cost. An interesting example of the working out of these problems can be seen in a groundbreaking regulation for heavy-duty engine emissions negotiated between the Agency and various interested parties. Where a small manufacturer finds the number of tests required by the Agency to establish a family's emissions level too burdensome, the firm may elect a sampling approach that uses fewer tests. The risk to the environment is held constant, leading to higher risk of having to pay unmerited non-compliance penalties in exchange for the smaller sample.

Distributed compliance systems that sounded wonderful when being discussed in theory by policy makers and economists may contribute to the development of ulcers by the Agency's legal fraternity. The very complexity of these schemes may become a major problem in court, where the violator can take pot-shots at the reasonableness of the regulation and seek refuge in the loopholes that are the unintended consequence of complexity. The statistical aspects of the design of the regulation are put to a severe test as the violator's attorneys and consultants question the Agency's proof that statistical assumptions were met or question the appropriateness of the methods chosen. Where compliance is distributed among different firms, major difficulties may arise over the fixing of responsibility for a violation--a problem that may be unlikely to occur with a simpler compliance scheme.

The case of lead

History and background

Lead compounds were first used in gasoline in the 1920s to boost octane. The effects of lead on octane can be seen in the sample response curve, Figure 1. While this curve is different for different base gasolines, its essential feature is a declining octane benefit per unit of lead as the total lead concentration increases. The nature of this curve creates an incentive for refiners to spread the amount of lead they are allowed to use as evenly as possible over the gallons of leaded gasoline produced. In addition to increasing octane rating, lead compounds provide some protection from valve wear to older engines designed with soft valve seats. This valve protection is provided by relatively low concentrations of lead compared to the more than two grams per leaded gallon (gplg) once used in leaded gasoline for octane reasons.

As mentioned earlier, lead in gasoline was first regulated in 1979 both to reduce lead for health reasons and to provide for availability of unleaded gasoline. Tougher standards for automotive emissions of carbon monoxide (CO) and hydrocarbons (HC) led auto makers to turn to catalytic converters as control devices. Widely used first in 1975, these devices are very sensitive to poisoning by lead, phosphorus, and other metallic substances.

Types of refineries

The refining industry grew up with the automobile and is thus a relatively old industry. Refineries are technologically stratified by age based upon the level of technology when they were constructed. The geographical development of the industry has tended to follow concentrations of population. Thus the older refineries tend to be located in the East. Newer refineries tend to be located near emerging centers of population and more recently developed sources of crude oil. These newer facilities, incorporating more recent technology, tend to be located on the West Coast.

As one might expect, refineries also vary considerably in size. Figure 2 shows something of the size distribution of the industry. A substantial number of these small refineries together produce only a small part of the total gasoline supply. In certain markets, these small facilities may play an important role due to high transportation costs from areas where larger and more efficient refineries are located.

The lead bubbles

Quarterly averaging. The first bubble or averaging approach used in regulating gasoline lead emerged almost unconsciously in the process of selecting an efficient way to monitor compliance. Since continuous monitoring of each refinery's output was not practical, and since requiring that each gallon of gasoline must meet a standard was very inflexible from the industry's standpoint, the first regulations prescribed a compliance period during which the average concentration of lead could not exceed the standard. The selection of a calendar quarter represents a compromise between environmental concerns and the industry's need for flexibility. The dimension for this bubble, then, is time. The relatively high concentrations dictate a short time span in order to protect public health. The integrating media are the air and soil from which lead emitted in automobile exhaust is taken into the human body. The environmental concerns regarding the use of the quarter are mitigated by the fact that the gasoline distribution system tends to mix gasoline from different producers in the marketplace, and the air and soil smooth out, over the course quarter, the intensity of human exposure.

Trading. The second bubble occurred in a more deliberate fashion with regulations that became effective in late 1982 and early 1983. These regulations shifted the basis of the standard and introduced a system of trading in lead usage rights. The standard was changed from one pertaining to a refinery's pooled gasoline output (unleaded and leaded considered together) to a standard applied strictly to leaded gasoline. The original regulation purposefully encouraged the increased production of unleaded gasoline as this product was new to the market. By 1982, unleaded gasoline had become a permanent fixture. The change to base the standard on leaded gasoline only was made so that the total amount of lead in gasoline would decline with the percentage of gasoline demand that was leaded. Under the older pooled standard the amount of lead per leaded gallon could increase as the percentage of leaded declined, resulting in a slower decline in total lead use.

Accompanied by a tightening of standards and a phaseout of special small refinery standards, the trading system provided for an improvement in the allocation of lead usage among refineries. This was done by permitting refineries which needed less lead than the standard allowed to sell their excess to other less technologically advanced refineries. Thus a modern facility capable of producing leaded gasoline comfortably at 0.70 gpig could sell the product of its leaded gallonage and the difference between that concentration and the standard of 1.10 gpig to one or more other refineries which found it necessary to use more than 1.10 gpig in their leaded gasoline. Such transactions were required to occur during the compliance period in question and could occur either within corporate boundaries or across them.

Without changing the time dimension, trading extended the bubble or distributed compliance system for lead into the dimension of space. Incurring no more transportation costs than the price of a stamp, a refinery or importer in New Jersey could purchase the right to use lead that was not needed by a refinery or importer in Oregon and thereby legitimize actual lead use that was over the standard. The integrating media were essentially the same as for quarterly averaging, but greater reliance was placed upon the homogenizing effects of the distribution system to avoid the development of "hot spots".

Banking. Responding to a mounting body of evidence on the negative health effects of lead and to the problem of increased conventional pollutants from lead-poisoned emission control systems, the Agency took further action on lead in early 1985. As shown in Figure 3, the resulting regulations reduced the allowable lead concentration by 91% in two stages (from 1.10 gpig to 0.50 gpig on July 1, 1985, and from 0.50 gpig to 0.10 gpig on January 1, 1986). This sharp tightening of the standard for lead was accompanied by a system of banking which effectively extended the lead bubble over a much longer time span than the calendar quarter that was previously allowed.

Under the banking provisions a refiner was allowed to store away in a bank account the difference between the standard and either 0.10 gpig or actual lead usage, whichever was larger. Such accumulation of rights was permitted during the four quarters of calendar 1985. The banked lead rights were to be available for use or transfer to another refiner or importer during any future quarter through 1987. Thus lead rights foregone during 1985 could be used to meet the sharply tighter 0.10 gpig standard during 1986 and 1987 after which any remaining rights expire. The 0.10 actual lead use limitation on rights accumulation was intended to avoid any incentive for refiners to use less than 0.10 gpig in leaded gasoline, since this was the level believed sufficient to protect the valves of some older engines from excessive wear.

The Agency's predictions of probable refiner behavior when given the flexibility of banking are shown in Figure 4, in which the concentrations from Figure 3 are weighted by estimates of leaded gallonage. The shaded areas during 1985 represent the extent to which Agency economists expected refineries to lower lead concentrations in order to bank lead rights for later use. The shaded areas farther to the right show the difference between the expected concentrations and the standard during the 1986-1987 period when the banked rights could be used to supplement the 0.10 gpig allowed under the standard. As the figure shows, the Agency expected only partial use of banking in the first quarter of 1985 due to the time required for refineries to revise their planning horizons under the new regulations. The heaviest banking was expected to occur in the second quarter as refineries were able to take full advantage of the regulation. The third and fourth quarters were expected to show only slight banking due to the 55% reduction in the standard to 0.50 gpig. Predictions for the 1986-1987 period show declining lead use in the second year as additional octane generation capacity was expected to come into service in anticipation of the 0.10 standard without banking.

This final step in extending a system of distributed compliance--a bubble--to cover lead in gasoline completed what was started by the decision to use quarters as compliance periods, greatly extending on a temporary basis the time span over which refineries could demonstrate compliance. Coupled with the trading provisions to provide for distribution over the space dimension, the package provided the industry with a very substantial degree of flexibility in meeting a standard which public health needs required to be as stringent as possible. The banking and trading together provided for an orderly adaptation by the more obsolete facilities, providing them with the time necessary to install new equipment.

How well it worked

Use of banking and trading. From the very beginning of the trading provisions in 1983, between one fifth and one third of the reporting facilities found it either necessary or desirable to purchase lead rights for use in demonstrating compliance with the regulations. The amount of lead involved in these transactions was at first small, amounting to about 7% of the total lead used. By the end of 1984 this figure had climbed to 20%.

The trading provisions of the regulation unintentionally permitted facilities blending alcohol into leaded gasoline to claim and sell lead rights based upon their activity. These facilities, frequently little more than large service stations, generated lead rights in the amount of the product of the 1.10 standard and the number of gallons of alcohol they blended. Both the lead and the gallons of leaded gasoline into which the alcohol was blended had already been reported by others. While these alcohol blenders increased sharply in number starting in the second quarter of 1984, their activities generated only a small amount of lead rights. This appearance of a new "industry" as an unexpected consequence of the regulation should remind the statistician or analyst that "ceteris paribus" is not always the case. Even with all the available information about the regulated industry to analyze, all else will not be equal since the regulation itself will cause perturbations, such as the new and previously nonexistent class of blender "refiners".

The banking program provided a great deal of flexibility to the industry, and accordingly was heavily used from its outset in the first quarter of 1985, even though the regulations were not made final until after the end of the quarter. About half of the entities reporting to the Agency made deposits in that first quarter, and the industry held the actual lead concentration to 0.70 gpig--lower than the Agency had predicted--thus banking more lead rights than expected. Along with the banking came a sharp increase in trading activity. The lead rights, because they no longer expired at the end of each quarter, were worth more and were traded in a more rational market where sellers had more time to seek out buyers and where brokers arose to place buyers and sellers in touch with each other. The higher price of lead rights led to an explosion in the number of alcohol blenders. Major refiners' facilities, which were previously not motivated to buy or even sell lead rights, began to bank and trade aggressively, stocking up rights for use in the 1986-1987 transition period at the new more stringent standard of 0.10 gpig.

Figures 5 and 6 show the lead use outcome of banking and trading compared to the standards and Agency predictions at the time the standards were promulgated. Figure 5 shows concentrations while figure 6 introduces leaded gallonage. The early and vigorous banking reduced concentrations to a lower level than expected, and substantial banking continued to occur on into the second half of the year under a half gram standard. Actual lead use, as figure 6 shows, was higher than predicted in both the second and third quarters as a result of higher than anticipated leaded gasoline usage. In all, 1985 ended with a net collective bank balance in excess of ten billion grams.

The first quarter of 1986 saw lead rights leaving the bank at about the rate that the Agency had predicted. The second quarter caused some alarm with a sharp drain on the bank owing to the unusually high leaded gallonage at a substantially higher concentration, 0.40 gpig, than predicted. As Figures 5 and 6 show, though, this early drain was partially offset by lower than expected usage in the fourth quarter.

The environmental effect of the regulation has been an unusually sharp and rapid decrease in a major pollutant, one that health studies indicate may be more dangerous at lower concentrations and to a broader segment of the human population than used to be believed. The banking and trading appear to have done precisely what they were intended and expected to, trading off lead use lower than the standard in 1985 against higher use in 1986-1987 with a total lead use over the period about the same as if the standards had been rigidly held to. It may be the case that a lead reduction this severe could not have been achieved without the distributed compliance approach that was used. It is certainly true that a transition to lower standards was achieved with greatly reduced economic impact.

Administration and enforcement. The banking and trading regulations were conceived with every intent that the Agency could keep a low profile and let market mechanisms do most of the work. While this was achieved to a substantial degree, the need to ensure compliance involved the Agency in processing more paperwork than the drafters of the regulations anticipated. It is probably worthwhile to examine briefly how this happened.

The flood of alcohol blenders swelling the ranks of the reporting population was not expected. Blenders had first come onto the scene with the trading provisions. By the end of 1984 they numbered something over a hundred, selling small amounts of lead credits, generated during the quarter, to small and/or obsolete refineries which were not otherwise able to meet the 1.10 gpig standard. In the first quarter of 1985 well over 200 additional blenders reported, drawn by the prospect of either immediately selling their lead usage rights at the sharply higher prices that prevailed with banking or retaining them and speculating on the price. As the word of this opportunity spread among distributors and service station chains, the population of these "refineries" exploded, reaching more than 600 by the third quarter of 1985 and pushing the reporting population above 900.

The numbers themselves would not have been such a problem for the Agency if all of the reports had been made correctly. The blenders, though, were new to this business. They didn't understand the regulations, and they lacked the accounting and legal departments which usually handled reporting for large refineries. The most common error made by the blenders was to attempt to bank and immediately sell to another refiner lead rights that could not legitimately be claimed. This frequently took the form of simply multiplying the alcohol gallonage by the standard (1.10 or 0.50 gpig, depending on the quarter), ignoring the restriction mentioned earlier that lead rights could be banked only on foregone lead usage above 0.10 gpig. By the time the blender filed a report and his error was detected by the Agency's computer, the rights had already been sold to another party and perhaps resold or used. In addition to the obvious legal tangle caused by this, there was the instability of the blender population--the party responsible for the improperly generated rights could not always be found.

The enforcement machinery developed by the Agency to handle lead phasedown was shaped by certain reasonable expectations about the reporting population--scale of operations, number of reporting entities, relative sophistication, etc. The blenders did not fit these expectations, and the enforcement process developed considerable congestion until some adaptation could take place. The computer system developed to audit reports and especially to match up the parties in lead rights transfers did precisely what it was designed to do and generated thick stacks of error output where only a few errors had been expected. The further processing of the errors had to be done manually and required clerical and legal staffing at a level that was not anticipated. By the time these resources were increased to the appropriate levels the backlog of errors was substantial and the time elapsed since the filing of the original reports made sorting things out more difficult.

A further illustration of how the crystal ball can fail is found in the difference between true refineries and the blenders in scale of operations. True refineries deal in such large quantities of gasoline and lead that for convenience all of the report forms used thousands of gallons and kilograms of lead as units. To report in smaller units would be to claim a degree of precision lacking in the basic information available to the refineries' accounting departments. The effect of rounding to thousands, trivial to larger refineries, was definitely not trivial to the blenders, many of whom only blended a thousand gallons of alcohol in a quarter. The blenders used whatever units optimized their profit with a fine disregard for the proper placement of decimal points. Where their gallonage was, say, 1,600 gallons, they would take advantage of the rounding instructions on the form to claim credits based upon 2 units of a thousand gallons each. If the amount was 1,400, they would report in gallons rather than thousands of gallons, often without labelling the units or putting a decimal point in the correct position.

All of these difficulties of enforcement logistics came into being as a result of the complexity of the bubble or distributed compliance system. With a simple set of rigid standards there would have been no blenders. Fortunately, this was a case where the environment suffered almost no harm as a result of the unforeseen consequences of the regulations, however embarrassing the situation may have been to Agency managers. This was probably mostly good luck, and should not be counted upon to happen routinely.

Legal Considerations

The statistician frequently finds himself with a well-thought-out concept for a procedure only to be faced with complications in the implementation scheme. Banking and trading proved no exception to this problem. The idea of free trade of lead rights between parties in order to increase flexibility of each refinery's planning was too good to resist. The government even took great pains to stay at "arms distance" in the trading process. Prior experience with the Department of Energy's entitlements program, in which the Federal government established formula upon formula to assure that every refinery got its "fair share" demonstrated that the Federal government was not the best broker in the refinery industry! In this case the EPA was staying out of the business.

So, what could go wrong? Since lead rights are valuable, there is an incentive to cheat. The value of lead rights rose from 3/4 of a penny to slightly over 4 cents per gram of lead. Trading and banking transactions are frequently in the order of 25 to 50 million grams. Thus the dollar amounts are in the $1 to $2 million dollar vicinity. Consequently, monitoring and enforcement become major issues. Monitoring and its requirement for extra personnel and computer usage has already been discussed. Enforcement and the legal considerations are another matter. Prior to banking and trading, the regulations were applied on a refinery by refinery basis and enforcement was a fairly straightforward matter. Under banking and trading the host of possible violations increased exponentially. The types of violations included trading rights that were improperly generated, selling the same rights twice, and banking rights for a future quarter that were in fact required for the current quarter's compliance. Any of these transgressions, of course, may have ramifications for the buyers of such lead rights. The situation becomes very complex from an enforcement standpoint since frequently rights are sold to an intermediary who resells them. If the original rights were bogus, or partly bogus, who among all the recipients has good rights and who has bad ones? These are not like counterfeit bills; they are entirely fungible, and determining if a particular right is legitimate can be a nightmare. Since banking lasts over several time periods, bogus rights can be exchanged frequently, and tracing the source of the bad rights can be next to impossible. Further, what action, if any, should be taken against the good faith purchaser of such lead rights? This last question subdivides into possible different actions depending upon whether the purchaser just deposits the rights into his account or, alternatively, actually uses them before they are discovered to be bogus. The possibilities seem endless!

An interesting sidelight to these difficulties is that it is frequently a small refiner with small amounts of rights that causes the difficulty. More effort is expended to chase small infractions than can be imagined, and enforcement policies designed for use with a small number of large violators prove awkward and unwieldy when dealing with a large number of small violators. A second side effect, though no fault of the designer of the regulation, is that many refineries find themselves bankrupt in today's oil industry. Chasing after lead rights of a bankrupt concern is generally far less than fruitful.

Nevertheless, the system has fared remarkably well. Over ten billion grams of lead rights were banked, roughly two year's worth, and no one is asking for government intervention to make lead rights trading run more smoothly. However, the point to be made is that the statistician can ill-afford to wash his hands of the problems involved in day-to-day implementation and enforcement of the requlations. He must guard against being the party who suggested the program and then walked away when some aspect didn't work as planned.

Conclusions

We have tried to provide in this paper an analytic framework for understanding the set of compliance management mechanisms loosely classified as "bubbles". We have seen something of the attractive features of such approaches, especially from the standpoint of the economic flexibility which they may make possible, but have also seen some of the ways in which things may go otherwise than as the drafters of the regulations intended. The lead phasedown banking and trading system was used to illustrate some of the concepts presented, even though the statistical problems in this regulation were less extensive than those with some other bubble regulations.

Distributed compliance schemes are fascinating to economists, and they are attractive to higher Agency managers from other professional backgrounds because of their potential to blunt the resistance to needed environmental regulation and sugarcoat the regulatory pill. The statistician must have a place in the development of these regulations--the questions of measurement, estimation, and uncertainty that are frequently involved demand it. The proper role of the statistician is not just that of picking up the pieces after things begin to go wrong in implementation. Neither is it to be a nit-picking nay-sayer whose business is to tell people why "you can't get there from here". Rather the statistician's role should be an affirmative one--that of a full partner in the regulation development process. As such, members of the profession must not only serve in the critical role of assuring a regulation's scientific integrity (and therefore its enforceability) but must also lend their creativity and special insights to the fundamental design of the regulation's compliance system, finding ways to do things where others, perhaps, cannot.

Figure 1
Gasoline octane enhancement from lead antiknock compounds

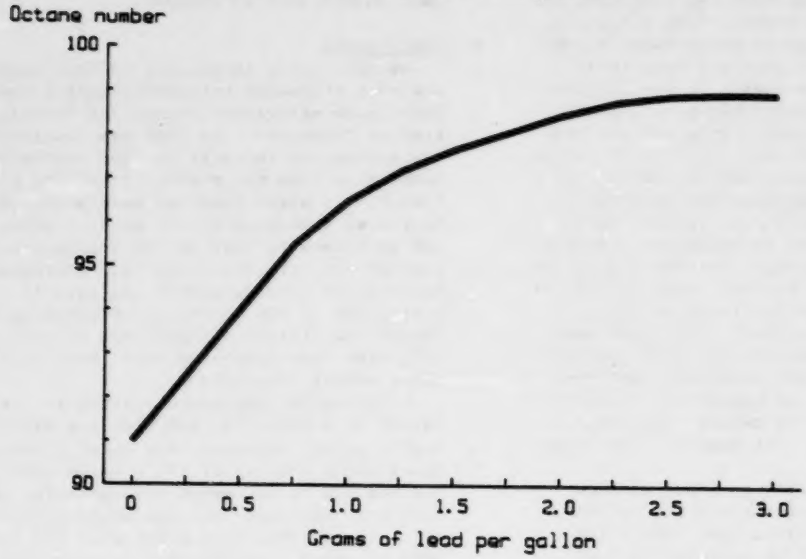
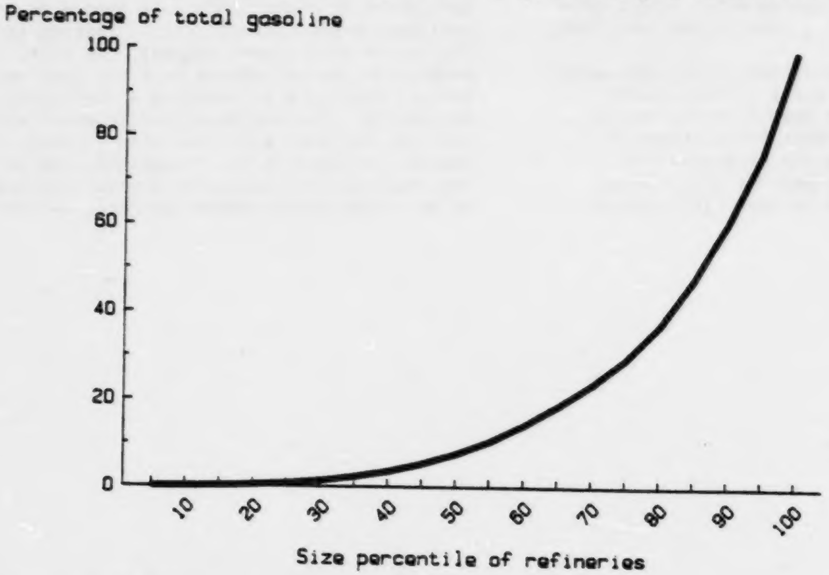
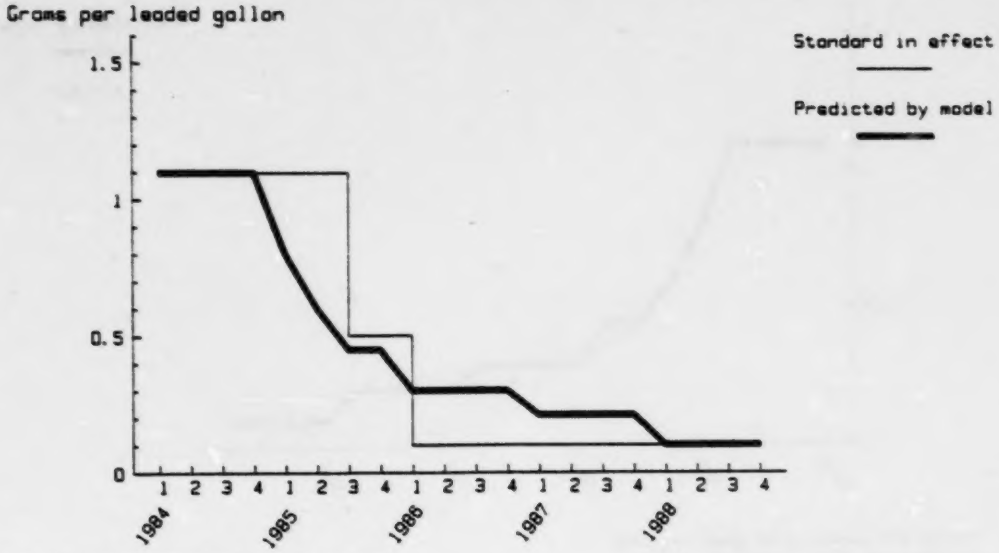


Figure 2
Cumulative percentage of total gasoline production by refinery size percentile*



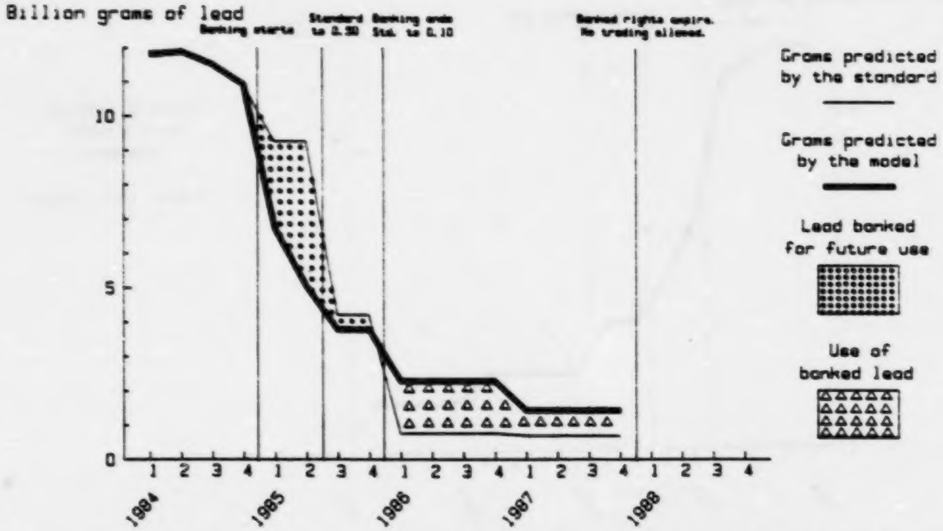
*Quarter III, 1983

Figure 3
Standards and predicted* lead concentrations under banking and trading



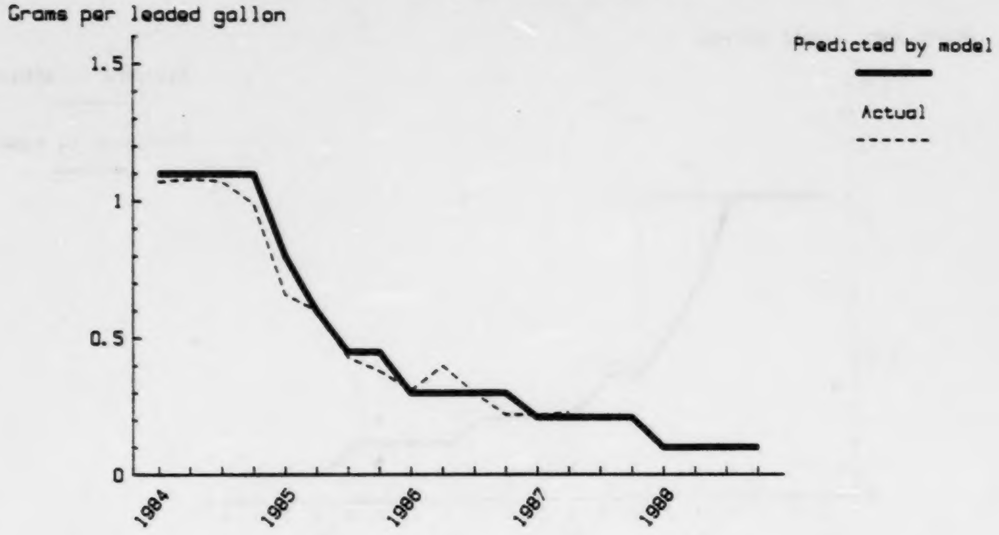
*Costs and Benefits of Reducing Lead in Gasoline, Feb., 1985, p. II-63.

Figure 4
Lead usage predicted with and without banking program*



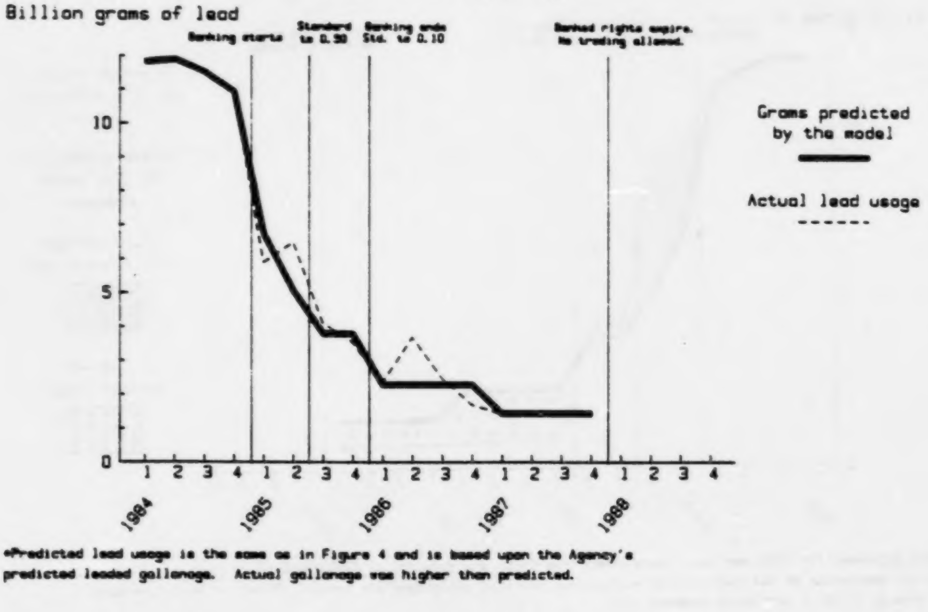
* Leaded gallonage for 1985 and later years taken from Costs and Benefits of Reducing Lead in Gasoline, Feb., 1985, with the assumption of 80% reduction in misfueling. Earlier gallonages are actual. Predicted concentrations are from p. II-63 of the above document.

Figure 5
Predicted* and actual lead concentrations
under banking program



*Costs and Benefits of Reducing Lead
in Gasoline, Feb., 1985, p. II-63.

Figure 6
Predicted and actual lead usage
with banking program*



*Predicted lead usage is the same as in Figure 4 and is based upon the Agency's predicted leaded gallonage. Actual gallonage was higher than predicted.

DISCUSSION

N. Phillip Ross
US Environmental Protection Agency

The concept of bubbles is intriguing; an umbrella under which trades can be made which enable regulated industries within the bubble to meet environmental standards--standards that they otherwise may not have been able to satisfy. This paper describes such a bubble; an umbrella of time for compliance with lead in gasoline standards.

The idea has logical appeal. Unfortunately, the world in which it is implemented is not always as logical. There is an implicit concept of uniformity that underlies the ideas of trading and banking. It's okay to have high levels of pollutants as long as you balance them against low levels either at a later point in time or by purchasing "credits." Although the "average" levels of the pollutant within the bubble's boundaries may be at or below the EPA standard, there will be many points within the bubble where levels are well above the standard. From a public health point of view, this may not be desirable. It eventually translates into periods when the population at risk will receive exposures to levels greater than the standard.

As pointed out by the authors, a major advantage to use of the bubble in the case of lead in gasoline was that many refiners and blenders who could not immediately meet the standard were able to continue operations through the purchase of credits. Indeed, imposition of the standard on many of these companies may well have forced them out of business. This is not a minor concern. Enforcement of environmental standards is exceptionally difficult. The regulated industry must be willing to cooperate through voluntary compliance. The bubble approach, even under conditions of non-uniformity, provides the needed incentives to encourage voluntary compliance. Environmental standards which cause major economic hardship for the regulated industry will be difficult to enforce. Federal enforcement resources are minimal. Lack of a substantial enforcement presence could result in greater pollution through noncompliance. Even though the real world does not always conform to the basic assumption of the bubble model, the real world will use the approach to achieve an overall reductions in pollution.

The lead bubble was very successful. As the authors have pointed out, there were problems; however, overall the levels of lead in gasoline did go down rapidly. This probably would not have happened under the more traditional approach to enforcement.

I agree with the author's conclusion that statisticians must learn to play a greater role in developing the strategies and in "finding ways to do things where others, perhaps, cannot." Statistical thinking involves the consideration of uncertainty in decisionmaking. All problems cannot be solved statistically; however, statistical thinking can help solve problems. Statisticians need to realize that their roles are not limited to the design or analysis components of a study. They have a role to play in the process of regulation development and in the development of new an innovative ways to deal with enforcement and compliance problems--ways which are not necessarily based on mathematically tractable assumptions.

VARIABLE SAMPLING SCHEDULES TO DETERMINE PM_{10} STATUS

Neil H. Frank and Thomas C. Curran
U.S. Environmental Protection Agency, Research Triangle Park, NC 27711

Introduction

In April 1971, EPA set National Ambient Air Quality Standards (NAAQS) for particulate matter (PM) and five other air pollutants - nitrogen dioxide, sulfur oxides, carbon dioxide, hydrocarbons, and photochemical oxidants.[1] There are two types of NAAQS: primary standards designed to protect human health and secondary standards designed to protect public welfare. In recent years, the standard for hydrocarbons has been rescinded and standards for an additional pollutant, lead, have been added. The reference method for measuring attainment of the PM standards promulgated in 1971 was the "high-volume" sampler, which collects PM up to a nominal size of 25 to 45 micrometers (um). This measure of PM was called "Total Suspended Particulate (TSP)" and was the indicator for the 1971 PM standards. The primary (health-related) standards set in 1971 for particulate matter (measured as TSP) were 260 ug/m^3, averaged over a period of 24-hours and not to be exceeded more than once per year, and 75 ug/m^3 annual geometric mean. The secondary (welfare-related) standard set in 1971 (measured as TSP) was 150 ug/m^3, averaged over a period of 24 hours and not to be exceeded more than once per year.

The gaseous NAAQS pollutants including carbon monoxide, nitrogen dioxide, ozone, and sulfur dioxide, are sampled with instruments which operate continuously, producing data for each hour of the year. This data is subsequently processed into various statistical indicators necessary to judge air quality status and attainment with their respective standards. Lead and TSP are NAAQS pollutants sampled on an intermittent basis. For these pollutants, one integrated 24-hour measurement is typically scheduled every sixth day. This is designed to produce measurements which are representative of every day of the week and season of the year. This approach has been shown to be useful in producing unbiased estimates of quarterly and annual average air quality, but has various limitations regarding estimation of peak air quality values. One shortcoming of concern was that attainment of the short-term 260 ug/m^3 TSP standard could be judged using data typically collected every sixth day and there was no specified adjustment for the effect of incomplete sampling. This was recognized as a problem in the early 1970's. If the second highest observed TSP measurement was less than 260 ug/m^3, the primary health related standard was judged as being attained. These standards were termed "deterministic."

Pursuant to the requirements of the 1977 amendments to the Clean Air Act, EPA has reviewed new scientific and technical data and has promulgated substantial revisions to the particulate matter standards.[2,3] The review identified the need to focus from larger, total particles to smaller, inhalable particles that are more damaging to human health. The TSP indicator for particulate matter has therefore, been replaced with a new indicator called PM_{10} that only includes those particles with an aerodynamic diameter smaller than or equal to a nominal 10 micrometers. A 24-hour concentration of 150 ug/m^3 levels was selected to provide a wide margin of safety against exposure which is associated with increased mortality and aggravation of respiratory illness; an annual average concentration of 50 ug/m^3 was selected to provide a reasonable margin of safety against long-term degradation in lung function. The secondary standards were set at the same levels to protect against welfare effects. The EPA review also noted that the relative protection provided by the previous short-term PM standards varied significantly with the frequency of sampling. This was identified as a flaw in both the form of the earlier TSP standard and the associated monitoring requirements. Following the recommendations of the EPA staff review, the interaction between the form of the standard and alternative monitoring requirements was considered in developing the recently promulgated PM standards.

Form of the New PM_{10} Standards

The new standards for particulate matter are stated in terms of a statistical form. The 24-hour standards were changed from a concentration level not to be exceeded more than once per year to a concentration level not to have more than one expected exceedance per year. This form corresponds to the one promulgated for the revised ozone standard in 1979.[4] The annual standards were changed from an annual average concentration not to be exceeded to an expected annual average concentration. To be more consistent with pollutant exposure, the annual average statistic was also changed from a geometric mean to an arithmetic mean.

The attainment tests, described for the new expected value forms of the particulate matter standards, are designed to reduce the effects of year-to-year variability in pollutant concentrations due to meteorology, and unusual events. For the new 24-hour PM standard, an expected annual number of exceedances would be estimated from observed data to account for the effects of incomplete sampling following the precedents set for the ozone standard. With averaging of annual arithmetic means and estimated exceedances over a multiple-year time period, the forms of these standards will permit more accurate indicators of air quality status and will provide a more stable target for control strategy development.

The adjustments for incomplete data and use of multi-year time periods are significant improvements in the interpretation of the particulate matter standards. These changes increase the relative importance of the 24-hour standard and play an important role in the development of the PM_{10} monitoring strategy. They also help to alleviate the implicit penalty under the old form that was associated with more complete data. The review of alternative forms of the 24-hour standards identified that the ability to detect nonattainment situations improves with increasing sample size. This is true for the previous "deterministic" form and the current statistical form. With the new 24-hour attainment test, however, there is a significant increase in the probability of failing the attainment test with incomplete data sets. This sets the stage for attainment sampling strategies.

Figure 1 presents the probability of failing the 24-hour attainment tests for the new PM_{10} NAAQS over a 3-year period. These failure probabilities were based on: (1) a constant 24-hour PM_{10} exceedance probability from an underlying concentration frequency distribution with a specified characteristic high value (concentration whose expected number of exceedances per year is exactly one), and (2) a binomial distribution of the number of observed exceedances as a function of sample size. Lognormal distributions with standard geometric deviations (sgd) of 1.4 and 1.6 were chosen for this illustration to represent typical air quality situations. The approach used in Figure 1 and throughout this paper are similar to analyses presented elsewhere.[5,6,7] This facilitates examining properties of the proposed standard in terms of the relative status of a site to the standard level (e.g. 20 percent above the standard or 10 percent below the standard) and the number of sampling days per year. It is worth noting that the percent above or below the standard is determined by the characteristic high. This is more indicative of the percent control requirements than using the expected exceedance rates.

Sampling frequency was judged to not be an important factor in the ability to identify nonattainment situations for either the current or previous annual standards. This is due to the generally unbiased nature and small statistical variability of the annual mean which is used to judge attainment with this standard. The change to an expected annual mean form, however, would tend to provide better estimates of the long-term pollutant behavior and provide a more stable indicator of attainment status.

With the new 24-hour attainment test, one important consequence of increased failure probabilities is the potential misclassification of true attainment areas. In Figure 1, it can be seen that these Type I errors are generally higher for small sample sizes, including those typical of previous TSP monitoring. This error is shown to be as high as 0.22 for a site which is 10 percent below the standard and has a sampling frequency of 115 days per year.

During the review of the standards, it was recognized that the ideal approach to evaluate air quality status would be to employ everyday sampling. This would minimize the potential misclassification error associated with the new PM attainment tests. From Figure 1, it can be seen that this would produce the desirable results of high failure probabilities for nonattainment sites and low failure probabilities for attainment sites. Unfortunately, existing PM monitoring technology as well as available monitoring resources do not make it convenient to monitor continuously throughout the nation. Moreover, while more data is better than less, it may not be necessary in all situations. When we revisit Figure 1, it can be seen that when a site is considerably above or below the standard, small sample sizes can also produce reasonably correct results with respect to attainment/nonattainment decisions. Thus, in order to balance the ideal and the practical, a monitoring strategy was developed which involves variable sampling schedules to determine PM_{10} status and attainment with the new standards.

The new strategy will permit most locations to continue sampling once in 6 days for particulate matter. Selected locations will be required to operate with systematic sampling schedules of once in 2 days or every day. With approval of EPA Regional Office these schedules may also vary quarterly depending on the local seasonal behavior of PM_{10}. Schedules of once in 3 days were not considered because of the discontinuity in failure probabilities occuring at 115 sampling days per year (95% data capture), seen in Figure 1 and discussed elsewhere.[5,7]

Monitoring Strategy

The previous monitoring regulations which applied to particulate matter specified that "at least one 24-hour sample (is required) every 6 days except during periods or seasons exempted by the Regional Administrator."[8] The new PM_{10} monitoring regulations would permit monitoring agencies to continue this sampling frequency for PM_{10} but would require them to conduct more frequent PM_{10} sampling in certain areas in order to estimate air quality indicators more accurately for control strategy development and to provide more correct attainment/nonattainment determinations.[9] The change in monitoring practice is largely required to overcome the deficiency of existing sampling frequency in detecting exceedances of the 24-hour standard. The operating schedules proposed for the measurement of PM_{10} will consist of a short-term and long-term monitoring plan. The short-term monitoring plan will be based on the requirements and time schedules set forth in the new PM_{10} Implementation Regulations for revising existing State Implementation Plans (SIPs).[10] The requirements ensure that the standards will be attained and properly maintained in a timely fashion. The long-term requirements will depend on PM_{10} air quality status derived from future PM_{10} monitoring data. These are designed to ensure that adequate information is produced to evaluate PM_{10} air quality status and to ensure that the standards are attained and subsequently maintained.

Consistent with the new reference sampling principle, available PM_{10} instruments only produce one integrated measurement during each 24-hour period. Multiple instruments operating with timers, therefore are necessary to avoid daily visits to a given location. The new standards, however, will permit approval of alternative "equivalent" methods which include the use of continuous analyzers. Because of the new monitoring requirements, instrument manufacturers are currently developing such analyzers. This will alleviate the temporary burden associated with more frequent monitoring.

Short-term Monitoring Plan

The proposed first-year monitoring requirements will be based on the requirements for revising SIPs. Areas of the country have been classified into three groups, based upon the likelihood that they are not currently attaining the PM_{10} standards as well as other considerations of SIP adequacy.[11] Since PM_{10} monitoring is in the process of being established nationwide and is quite limited, a procedure was used which estimated the probability that each area of the country would not attain the new standards using existing TSP data in combination with available PM_{10} data. This is described elsewhere.[12]

Areas have been classified as Group I, II or III. Group I areas have been judged to have a high probability, $p \geq 0.95$, of not being in attainment with the new standards. Group II areas have been judged to be too close to call, but still very likely to violate the new standards ($0.20 \leq p < 0.95$). Group III areas have been judged to be in attainment ($p < 0.20$).

For Group I areas, the value of a first year intensified PM_{10} data collection is most important. This is because these areas are most likely to require a revised SIP. Since the 24-hour standard is expected to be controlling, the development of control strategies will require at least 1 complete year of representative data. Consequently, everyday sampling for a minimum of 1 year is required for the worst site in these areas in order to confirm a probable nonattainment status, and to determine the degree of the problem.

The Group II category identifies areas which may be nonattainment (but whose air quality status is essentially too close to call.) For such areas, the value of additional PM_{10} information is important in order to properly categorize air quality status. For these areas, more intensified sampling is desirable. Based on the consideration of cost, and available monitoring resources, however, a more practical strategy of sampling once in 2 days at the worst site is required for the first year of monitoring.

All remaining areas in the country (defined in terms of $p<0.20$) have been categorized Group III and judged not likely to violate the new standards. For such areas, the value of collecting more than a minimum amount of PM_{10} data is relatively low and intensified PM_{10} data collection is not warranted. Recognizing that there is still a small chance of being nonattainment, however, a minimum sampling program is still required at these locations. Based on considerations of failing the 24-hour attainment test and estimating an annual mean value, a minimum sampling frequency of once in 6 days is required.

The short-term strategy also contains previsions for monitoring to be intensified to everyday at the site of expected maximum concentration if exceeddances of the 24-hour standard are measured during the first year of monitoring. This is intended to reduce the potential for nonattainment misclassification (type I error) with the 24-hour PM_{10} attainment test. With this provision, the first observed exceedance is not adjusted for incomplete sampling and is assumed to be the only true exceedance at that location during the calendar quarter in which it occurred. The effect on misclassification error associated with a 3-year attainment test is illustrated in Figure 2. It can be seen that the sites most vulnerable to this error are slightly less than the standard. In these comparisons, for sites which are 10 percent less than the standard and are sampling once in 2 days, the type I error is reduced from 6 percent to 1 percent. If these same sites are sampling once in 6 days, the type I error is similarly reduced from 12 percent to 0.5 percent. There is, however, a corresponding increase in the type II error associated with the attainment test for true nonattainment sites also close to the standard. This compromise was judged to be appropriate in developing the new rules.

Long-term Monitoring Plan

The long-term monitoring plan starts with the second year of sampling. The required sampling frequencies are based on an analysis of the ratio of measured PM_{10} concentrations to the controlling PM_{10} standard. This determination depends upon an assessment of (1) whether the annual or 24-hour standard is controlling and, if it is the latter, (2) the magnitude of the 24-hour PM_{10} problem. Both items are evaluated in terms of the air quality statistic called the design concentration. For the annual standard, the design concentration is the expected annual mean; for the 24-hour standard, the design concentration is the characteristic high value whose expected exceedance rate is once per year. In both cases the design concentration is the value the control strategy must be capable of reducing to the level of the standard in order to achieve attainment. The ratio to the standard is defined in terms of the design concentrations and the standard level; the controlling standard is simply the standard which has the highest ratio. This is a somewhat simplified definition but is adequate for present purposes.

The long-term strategy specifies frequencies of every day, every other day, or every sixth day. The long-term monitoring strategy is designed to optimize monitoring resources and maximize information concerning attainment status. As with the short-term strategy, the increased sampling frequency provisions only apply to the site with expected maximum concentration in each monitoring area.

For those areas where the annual standard is controlling, 1 in 6 day monitoring would be required; this frequency has been judged to be adequate for assessing status with respect to this standard. For those areas where the 24-hour standard is controlling, the required minimum sampling frequency for the calendar year will vary according to the relative level of the most current maximum concentration site to the level of the standard. In other words, the sampling requirement applies to the site which drives attainment/nonattainment status for the monitoring area. The least frequent monitoring (1 in 6 days) would be required for those areas where the maximum concentration site is clearly above the standard (≥40 percent above) or clearly below the standard (>20 percent below). For such sites a minimum amount of data collection would be adequate to verify correct attainment/nonattainment status. As the area approaches the standard, the monitoring frequency for the maximum concentration site would increase so that the misclassification of correct attainment/nonattainment status can be reduced. If the area is either 10-20 percent below or 20-40 percent above the 24-hour standard, 1 in 2 day monitoring would be required. When the area is close to the standard, i.e. 10 percent below to 20 percent above, everyday sampling would be required in order to improve the stability of the attainment/nonattainment classification. Figures 2 and 3 illustrate misclassification rates for a 3-year, 24-hour attainment test as a function of the relative status of a site to the standard and in terms of alternative sampling frequencies. As with previous analyses, underlying lognormal distributions with sgd's of 1.4 and 1.6 for attainment and nonattainment sites are utilized. For sites following the long-term incomplete sampling schedules (1 in 6 days and 1 in 2 days) misclassification rates can be maintained in or below the neighborhood of 5-10 percent.

Summary

The revisions to the PM standards improve the ability to identify nonattainment situations, provide for more stable pollutant indicators, and change the relative importance of the annual and 24-hour averaging times. With the required adjustments for incomplete sampling in the interpretation of PM data, the revised standard would correct for the variable protection afforded by the current 24-hour PM standard, and it is expected that the revised 24-hour standard will generally be controlling.

Monitoring requirements have been promulgated which will similarly correct for the deficiency in the current standards. Variable frequencies are now required in order to reduce the uncertainty associated with attainment/nonattainment classification. This provides more uniform protection by the standards but at the same time conserves scarce monitoring resources. The initial requirements will place the most emphasis on areas with the highest estimated probability of violating the PM_{10} standards while the long-term strategy will allow sampling frequency to vary according to the relative status of an area with respect to the standard concentration levels.

The operational difficulties associated with implementing the new requirements for everyday monitoring has generated new research initiatives to develop a continuous analyzer for PM_{10}. Once this is available, particulate matter can be conveniently monitored everywhere on the same basis as the gaseous NAAQS pollutants.

References

1. "National Primary and Secondary Ambient Air Quality Standards," Federal Register, 36(84):8186. April 30, 1971.

2. Review of the National Ambient Air Quality Standards for Particulate Matter: Assessment of Scientific and Technical Information, OAQPS Staff Paper. U. S. Environmental Protection Agency, Research Triangle Park, N.C. 27711. EPA-450/5-82-001. January 1982.

3. "Revisions to the National Ambient Air Quality Standards for Particulate Matter," Federal Register, 52(126):24634. July 1, 1987.

4. "Revisions to the National Ambient Air Quality Standard for Photochemical Oxidants," Federal Register, 44(28):8202. February 8, 1979.

5. Frank, N. H. and T. C. Curran, "Statistical Aspects of a 24-hour National Ambient Air Quality Standard for Particulate Matter," presented at the 75th APCA Annual Meeting, New Orleans, LA. June 1982.

6. Davidson, J. E., and P. K. Hopke, "Implications of Incomplete Sampling on a Statistical Form of the Ambient Air Quality Standard for Particulate Matter," Environmental Science and Technology, 18(8), 1984.

7. Frank, N. H., S. F. Sleva and N. J. Berg, Jr. "Revising the National Ambient Air Quality Standards for Particulate Matter - A Selective Sampling Monitoring Strategy," presented at the 77th Annual Meeting of the Air Pollution Control Association, San Francisco, CA., June 1984.

8. "Ambient Air Quality Surveillance," Federal Register, 44(92):27571, May 10, 1979.

9. "Ambient Air Quality Surveillance for Particulate Matter," Federal Register, 52(126):24736. July 1, 1987.

10. "Regulations for Implementing Revised Particulate Matter Standards," Federal Register, 52(126):24672. July 1, 1987.

11. "PM_{10} Group I and Group II Areas" Federal Register, 52(152):29383. August 7, 1987.

12. Pace, T. G., and N. H. Frank, "Procedures for Estimating Probability of Nonattainment of a PM_{10} NAAQS Using Total Suspended Particulate or Inhalable Particulate Data," U. S. Environmental Protection Agency, Research Triangle Park, N.C. 1984.

FIGURE 1. FAILURE PROBABILITIES FOR 3-YEAR, 24-HOUR ATTAINMENT TEST
WITH CONSTANT SAMPLING RATE

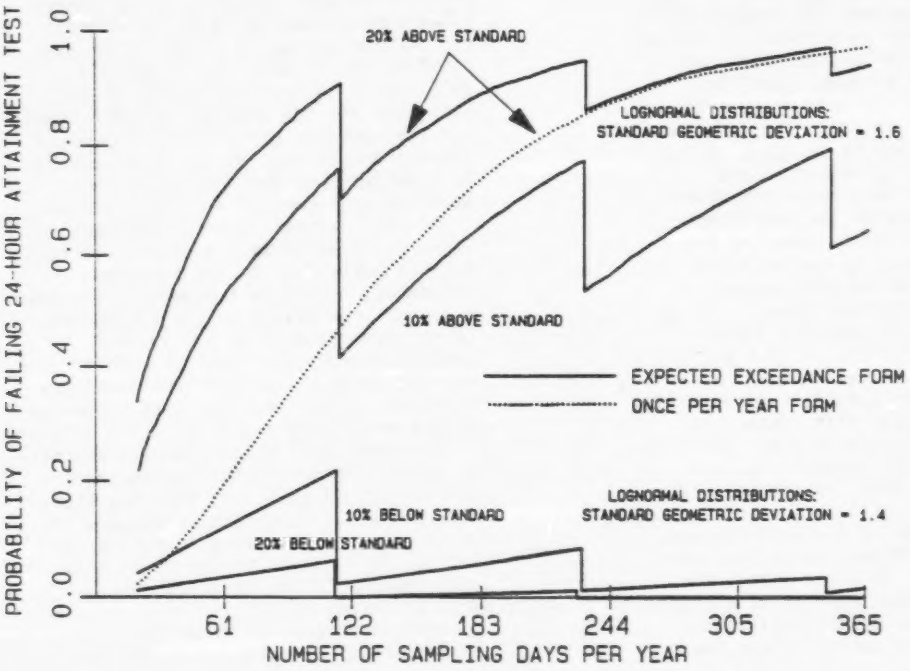


FIGURE 2. PROBABILITY OF NONATTAINMENT MISCLASSIFICATION

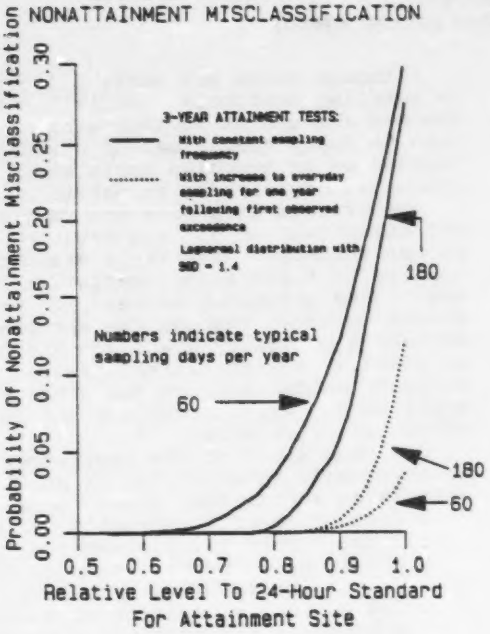
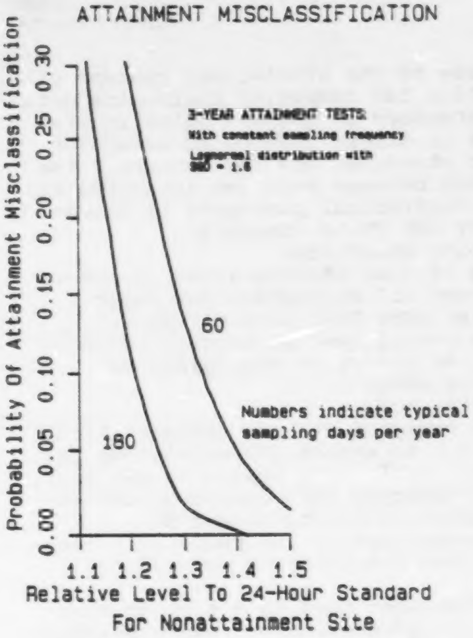


FIGURE 3. PROBABILITY OF ATTAINMENT MISCLASSIFICATION



DISCUSSION

John Warren
US Environmental Protection Agency

The use of the statistical concept of expectation for comparing monitoring data with a standard is new and quite intriguing as it offers promise of extension to other standards and regulations. The difference between existing standards and the new statistical standards is illustrated by the PM-10 standards.

Existing standards:

- o The 24-hour concentration is not to exceed 150 micrograms per cubic meter more than once per year.
- o The annual average concentration is not to exceed 50 micrograms per cubic meter.

New standards:

- o The expected 24-hour concentration is not to exceed 150 micrograms per cubic meter more than once per year.
- o The expected annual average concentration is not to exceed 50.

The advantages of the "expected" methodology over the existing methodology include:

- o It has been used in a similar fashion in generating the Ozone standard and therefore "familiar" to the public.
- o It uses actual data to generate the results.
- o There is a reduction in year-to-year variability.
- o It enables the development of stable control strategy targets.

The difference between the two methodologies would therefore appear to be small and hence readily adaptable to other standards. One possible candidate for the new methodology would seem to be Effluent Guidelines and Standards, Subchapter N, 40 CFR 400-471. These regulations stem from the Clean Water Act (1972) and are based on the engineering standards of Best Practicable Technology (BPT) or Best Available Technology (BAT). These guidelines cover mining industries (minerals, iron ore, coal etc.), natural products (timber, pulp and paper, leather tanning etc.), and the manufacturing industries (pharmaceutical, rubber, plastics, etc.). A typical standard within these guidelines is the Steam Electric Power Generating Point Source Category (Part 423.12, Effluent Limitations Using BPT):

BPT Effluent Limitations

Pollutant or Property	Maximum for any 1 day	Avgs. of Daily values for 30 consecutive days shall not exceed
Total Suspended Solids..........	100.0 mg/l	30.0 mg/l
Oil and Grease..	20.0 mg/l	15.0 mg/l
Copper, total...	1.0 mg/l	1.0 mg/l
Iron, total.....	1.0 mg/l	1.0 mg/l

Although there are small differences in sampling protocols, comparison with the new and old PM-10 standards would seem to imply that a set of standards devised on an expected basis would be possible; however, it is not to be.

The problem lies with the very different objectives of the regulations, state versus industry. The PM-10 standard applies to a State Implementation Plan, a negotiated agreement between EPA and the states enforced through the National Ambient Air Quality Standards and used to identify non-attainment areas. The Effluent Guidelines, on the other hand, apply to a specific industry and is not a matter of negotiation.

The resolution of the regulatory problems will be as difficult as the associated statistical problems of:

- o Assumption of lognormality of data
- o Stability of the process over time
- o Potential autocorrelation of data
- o Uncertainties of data quality
- o The optimal allocation of monitoring systems in non-attainment areas.

Despite these problems, it is clear that a statistical approach, in this case expected values based on an underlying lognormal distribution, is probably the way of the future; research should be encouraged in this field. Neil Frank and Thomas Curran have indicated a viable approach; where will the next step lead?

ANALYSIS OF THE RELATIONSHIP BETWEEN MAXIMUM AND AVERAGE IN SO2 TIME SERIES

Thomas Hammerstrom and Ronald E. Wyzga

1. Introduction and Motivation

Several studies have examined the physiological and symptomatic responses of individuals to various air pollutants under controlled conditions. Exposures in these experiments are often of limited duration. These studies demonstrate response with exposures as short as five minutes.

On the other hand, monitoring data rarely exist for periods as short as five minutes. Some measurement methods do not lend themselves to short term measurements; for other methods, 5-minute data often are collected but are not saved or reported because of the massive effort that would be required. In general, the shortest time average reported with monitoring data is one hour, and for some pollutants even this time average is too short.

Where monitored data do not exist, ambient concentrations can be estimated by the use of atmospheric dispersion models. The accuracy of these models degrades as averaging times decrease and they require meteorological and atmospheric inputs for the same time average as predicted by the model. Thus, air dispersion models are rarely used for time averages less than an hour.

There is, thus, a fundamental mismatch in time periods between health response and exposure, with responses occurring after only 5 or 10 minutes of exposure while exposure data are only available for periods of an hour or more. This paper attempts to address this mismatch by examining the relationship between a short-term time average (5 minutes) and a longer term time average (60 minutes) for one pollutant (SO2) for which some data are available. Understanding the relationship between the two time averages would allow the estimation of response given longer term estimates of ambient concentration. It could also help in the setting of standards for long term averages which would help protect against peak exposures.

This paper explores the type of inferences that can be made about five minute SO2 levels, given information on hourly levels. There are three possible models for health effects which motivate these inferences:

1. there is one effect in an hour if any 5-minute exposure level exceeds a threshold,
2. each 5-minute segment corresponds to an independent Bernoulli trial with probability of an effect equal to some increasing function of the current 5-minute level,
3. each 5-minute segment is a Bernoulli trial with the probability of an effect depending on the entire recent history of the SO2 process.

Corresponding to these health models, there are three possible parameters to estimate:

1. the distribution of the maximum 5-minute level during an hour,
2. the distribution of an arbitrary 5-minute reading,
3. the joint distribution of all twelve 5-minute readings.

All three distributions are conditional distributions, given the average of all twelve 5-minute readings. The first conditional distribution is the parameter of interest if one postulates that the dose response function for health effect is an indicator function and only one health event per hour is possible; the second is the parameter of interest if one postulates a continuous dose response function with each 5-minute segment constituting an independent Bernoulli trial; the third conditional distribution is of interest if one postulates that the occurrence of a health effect within an hour depends continuously on the cumulative number of 5-minute peaks.

This paper discusses some approaches to each of these three estimation problems. Section 2 discusses why the problem is not amenable to solution by routine algebra. Sections 3 and 4 present results for the estimation of the maximum. Section 3 presents some ad hoc methods for modelling the maximum as a simple function of the average when both are known and discuss how to extend these methods to estimate the maximum when it is unknown. Section 4 discusses the error characteristics of these methods. Section 5 presents an ad hoc method of estimating an arbitrary 5-minute level from the hourly average; Section 6 discusses the error characteristics of this method. Finally, Section 7 presents an estimation of the joint distribution of all twelve 5-minute readings, derived from a specific distribution--theoretic model for the 5-minute time series and discuss some of the difficulties involved with extending this.

2. Obstacles to Theoretical Analysis

A brief discussion of why we resorted to ad hoc methods is needed to begin with. In theory, given a model for the (unconditional) joint distribution of the time-series of 5-minute readings, it is straightforward to write down the exact formula for the joint conditional distribution of the twelve 5-minute readings, given the average.

If $\underset{\sim}{X} = (X_1, \ldots, X_p)$ has joint density $f(\underset{\sim}{x})$ and if $\overline{X} = \sum X_i/p$

then the conditional joint density is given by equation 1.

$$(1) \quad h(\underset{\sim}{x} \,|\, \overline{x}) = f(\underset{\sim}{x})\, I(\textstyle\sum x_i/p = \overline{x}) \Big/ \int_S f(\underset{\sim}{x})\, d\underset{\sim}{x}$$

where S is the simplex $\{\underset{\sim}{x} : \sum x_i/p = \overline{x}\}$ and I is the indicator function.

The conditional distribution of the maximum and the conditional distribution of any 5-minute reading would follow immediately from the conditional joint distribution of all twelve 5-minute levels.

Unfortunately, estimation of the unconditional joint distribution, f(x), of the 5-minute time-series is not easy. Non-parametric density estimation requires gigantic data sets when one is working in several dimensions.

Parametric modelling also poses formidable computational problems. If $f(x;\theta)$ is the joint density of the 5-minute levels, then the log likelihood function, based on observing only a sequence of N hourly averages, $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N$, is given by

$$(2) \quad L_N(\theta) = \sum_{1}^{N} \log \int_{S_i} f(\underset{\sim}{w};\theta)\, d\underset{\sim}{w} \, .$$

Here $S_i = \{\underset{\sim}{w} : \sum_{1}^{p} w_j/p = \overline{x}_i\}$ for i=1, 2, ... N.

Each term on the right hand side is the integral of a 12-dimensional density over an 11-dimensional simplex. For most reasonable choices of a joint distribution of the 5-minute readings, these integrals can only be evaluated numerically, using Monte Carlo methods. To find maximum likelihood estimates of θ, one must numerically evaluate L_N at sufficiently many values of θ to approximate the maximizing value. θ is always at least three dimensional (location, scale, correlation) and N will be in the hundreds (or thousands), making numerical maximum likelihood estimation a nearly insurmountable task. (Moreover, the hourly averages in the observed must not be consecutive hours but must be far enough apart in time to be effectively independent; otherwise, the likelihood function is even more complicated.)

An additional problem with parametric modelling is the choice of the functional form of the joint density f. One can test hypotheses that the hourly averages come from one of the commonly used distributions: lognormal, Weibull, or gamma. However, if hourly average SO2 readings are, say, lognormal, then 5-minute averages are not lognormal. In general, one would expect the hourly averages to be closer in shape to the normal distribution than are the 5-minute levels. (At least, this would be true if the 5-minute levels have the same finite variance.) There is no technique for inferring the functional form of the distribution of the individual terms in a sum from the functional form of the distribution of the sum.

As an alternative to theoretical modelling of the relevant conditional distributions, we have explored some ad hoc empirical methods of estimation. It is important to bear in mind that the objective of the exercise is not merely to determine a functional form for the relationship between 5-minute levels and hourly averages; but rather it is to provide specific numeric estimates that can be used when the five-minute levels are not observed. There are no unknowns when the five-minute levels are known so the only application of such a technique is extrapolation to situations where no data for new parameter estimation are available.

3. Estimation of the Maximum

3.1 Nature of the Data

The Electric Power Research Institute has collected data relevant to this inference from two different studies. The first comes from a group of stations monitoring a point-source; the second from a station monitoring ambient levels in a populated area. At these two sites, data were collected in each 5 minute segment for long periods of time, permitting direct comparison of the hourly and 5-minute levels. The first data set analyzed was from 18 monitors around the Kincaid power plant in Illinois, a coal-fired plant in Christian County, Ill., with a single 615 foot stack and a generating capacity of 1320 megawatts. The data set consists of nine months of observations from 18 stations around this plant. SO2 readings at these stations reflect the behavior of the plume from the stack. For a given monitor there are long stretches where SO2 levels are zero, indicating that the plume is not blowing toward the monitor. Such readings constitute about 72% of the hours in the data set; these were discarded before any further analysis was done. The second data set consists of SO2 data from a New York City monitoring station not near any dominant point source. The data were collected between December 15, 1981, and March 11, 1984.

3.2 Outline of Methods Used

We explored three empirical methods of estimating the maximum 5-minute reading from the hourly average. All three methods postulate a simple parametric model for the maximum as a function of the average. The methods differ only in how estimates of the parameters are obtained. The first method obtains parameter estimates from data containing 5-minute readings and then uses these estimates for other data sets collected elsewhere (and containing only 1 hour readings). This method is motivated by the theory that there is a universal law governing the relationship between the maximum and the average of an SO2 time series, with the same parameters at all sites. The second method requires expending effort to collect 5-minute data for a short period of time at the site of interest and using the data from this period to obtain parameter estimates that will be used over much longer periods when sampling is only on the 1-hour basis. The third method fits a simple parametric model to the maximum hourly reading in a 12-hour block as a function of the average over the 12-hour block and then assumes that the same model with the same numeric estimates describes the maximum 5-minute level in an hour as a function of the hourly average. (Daily cycles are removed from the 12-hour block data prior to estimation by dividing by long-term averages over a fixed hour of the clock.) For mnemonic purposes, we will call these three methods: 1. the method of universal constants, 2. the method of short-term monitors, and 3. the method of change of time-scale.

Estimates of the potential errors in the method of universal constants were obtained by using the parameter estimates from the New York data to fit the Kincaid data and vice versa. Potential errors in the method of short-term monitors were estimated by dividing both data sets into batches 100 hours long and then using each of the hundred odd resulting parameter estimates to fit 13 randomly selected hours. The hours were chosen by dividing the range of hourly averages into 13 intervals and choosing one hour from each interval. Potential errors in the method of change of time-scale were obtained by simply comparing the maxima predicted using the estimates from the 12-hour blocks in each data set with the observed maxima in the same data.

3.3 Parametric Models for the Maximum

The parametric models proposed here are intended to give ad hoc approximations to the maximum. One can show that they cannot be the true theoretical formulae. Because the maximum necessarily increases as the average increases, it is more convenient to work with the ratio of the maximum to the average than with the maximum itself. Previous authors (Larsen et al., 1971) working on this problem have used models in which log(ratio) is linear in log(average). Therefore, we began by fitting such a model to the two data sets by ordinary least squares. These estimates are given in Table 1. As may readily be checked, for both data sets, this model leads to impossible values, fitted ratios which are less than one, for large values of the average. For the Kincaid data, this occurs at relatively low values of the average.

In fact, it is not thought that a single universal set of constants applies to the regression of log (ratio) on log(average). Rather, it is thought that the atmospheric conditions around the monitor are classified into one of seven stability classes; and it may be more appropriate to assume the parameters of the regression are constant within a given a stability class. It is possible that the impossible values of the fitted maximum occur because of a Simpson's paradox in the pooling of data from several stability classes. Ideally, the above model should be fitted separately to each stability class. Unfortunately, there were no meteorological data available to permit such a partition of the data. It is possible that it would be worthwhile to obtain such data and redo the analysis. The difference between the Kincaid and New York City sites must be emphasized. The sources and variability of pollution are very different, and it may not be reasonable to extrapolate from one site to another; two data sets from like sites should be considered in subsequent analyses.

In order to prevent the occurrence of impossible fitted values, we fit models in which the log[log(ratio)] is a linear function of the log(average). The ordinary least square (OLS) estimates (for New York and Kincaid) of this line are also given in Table 1. Figures 1 and 2 show the scatter plots of the maximum vs the average. Both axes have logarithmic scales. If the log of the ratio were linear in the log of the average, one would expect that the vertical width of the scatterplot would remain roughly constant as the average varied. Instead, it appears that the scatterplots narrow vertically as the average increases, as would be expected if the iterated logarithm of the ratio were linear in the log of the average. For both data sets, it appears that the iterated log log model more accurately mimics the real data than the only former model shows the diminishing (on log scale) spread of the maximum with increasing values of the hourly average. This model is the preferable one to estimate the maximum.

In both data sets, the residuals were slightly negatively skewed with the skewness being greater in the Kincaid data. It seems reasonable to assume that the residuals in the New York data were approximately normal. This assumption is harder to maintain for the Kincaid data. Figures 3 and 4 show the histograms and normal probability plots for the residuals from these two regressions.

The main purpose of the analysis is to obtain a formula for estimating the conditional distribution of the unobserved 5-minute maxima from the observed hourly averages. The iterated log vs log models yield the following two formulae, given in equations 2 and 3.

(2) Prob(5-minute max$\le$ x | hourly average = y) =

$$F(x|y) = \Phi(\{\log\log(x/y)+.267*\log(y)+.719\}/.62\ )$$

for New York

(3) Prob(5-minute max$\le$ x | hourly average = y) =

$$F(x|y) = G(\{\log\log(x/y)+.258*\log(y)+.191\}\ )$$

for Kincaid.

Here Φ is the normal cumulative distribution and G is the empirical distribution function of the residuals of the OLS regression of loglog ratio on log average. We recommend using G in place of treating these residuals as normal. G is tabulated in table 2; its histogram is graphed in figure 4. Equations 2 and 3 do a reasonably good job of modelling the observed maxima in the two data sets from which the values of the parameter estimates were derived.

Inverting equations 2 and 3 gives simple formulae for the percentiles of the conditional distribution of the 5-minute maxima. Notice that equation 3, table 2, and linear interpolation permit estimation of percentiles of the Kincaid maxima from the 5'th to the 95'th. Attempts to estimate more extreme percentiles would require foolishly rash extrapolation.

The log vs log models provide a competing (and somewhat inferior) method of estimation. They yield conditional distributions of the 5-minute maxima given by equations 4 and 5.

(4) Prob(5-minute max$\le$ x | hourly average = y) =

$$F(x|y) = \Phi(\{\log(x/y)+.077*\log(y)-.499\}/.2\ )$$

for New York

(5) Prob(5-minute max$\le$ x | hourly average = y) =

$$F(x|y) = \Phi(\{\log(x/y)+.21*\log(y)-1.07\}/.69\ )$$

for Kincaid.

In these regressions, we found it acceptable to use a normal approximation for the residuals in both New York and Kincaid.

4. Error Estimation

4.1 Errors in the Method of Universal Constants.

It is not feasible to use a conventional method to estimate the uncertainty in the maxima fitted with this method. The major difficulty is that one is not looking for a well-behaved estimator but rather for a particular numeric value of the estimate for use in all data sets. The standard error of the estimate in one data set is quite misleading as a measure of the error that would result from using that same estimate in another data set. A further exacerbation results from the high correlation between the observations used to generate the estimates. The conventional formulae for the standard errors will exaggerate the amount of information in the data set and yield spuriously small standard errors. Finally, there is the problem that one knows that the model is theoretically incorrect and that the true underlying distribution is unknown so the conventional standard error formulae based on the modeled distribution are necessarily in error. One would suspect that even if the model adequately approximates the first moment of the maximum, it approximates the second moment less well.

As an alternative method for estimating the uncertainty in the method of universal constants for all data sets, a cross-validation method was pursued. We used the estimated parameters from each of the New York and Kincaid data sets to estimate the maxima for the other data set. For each hour, the estimated maximum were divided by the actual maximum, the resulting ratios were grouped into 10 bins, according to the value of the hourly average. Within each of these bins, we computed the three quartiles of the quotients of fitted over actual maxima. Figures 5 and 6 show these three quartiles of the fitted over true ratios, plotted against the midpoint of the hourly averages in the bin.

One should recall that the Kincaid data reflect the situation near a point source while the New York data reflects ambient levels far from any point source. Consequently, this method of cross-validation may exaggerate the error associated with this procedure. However, unless additional 5-minute data are collected and analyzed from a second plant and from a second population center station, it is difficult to determine how much of the error is due to the disparity of sites and how much due to the method.

The most striking feature of these plots is that the two cross-validations are biased (necessarily, in opposite directions). The higher values of the hourly average (the right half of the graph) are of greater interest. For the New York data, the first quartile of the ratio of fitted over the actual maximum is greater than 1; i.e. the estimated maximum is too high three fourths of the time. The median of the fitted over actual ratio is, for most hourly averages over 1.2; i.e. the estimated maximum is 20% too high more than half the time. The estimated maximum is 30-40% too high at least a quarter of the time. The situation at Kincaid is essentially the mirror image of this: for the higher values of the hourly average, the third quartile of the fitted over actual ratio is below .9; i.e. estimated maxima are at least 10% too low nearly three fourths of the time. They are 30-40% too low nearly half the time; are 50-60% too low at least a quarter of the time.

The proportionate error diminishes as the hourly average goes up. This, of course, is an artifact of using fitted value/true value as the measure of error. In absolute size (ug/m^3), the errors would not diminish as the hourly average increases.

4.2 Errors in the Method of Short-term Monitors.

In order to estimate the errors associated with attempting to estimate parameters of the ratio-average relationship at a given site by actually measuring 5-minute levels for a short time, each data set was divided into batches 100 hours long and OLS estimates were derived for each batch. There are 125 such batches in the New York data and 158 batches in the Kincaid data.

It is difficult to judge the potential in estimating the maxima by simply looking at the uncertainty in these parameters. In order to further clarify the errors of direct interest, we divided the hours into 13 bins, according to the size of their hourly averages. For each OLS estimate from a batch, we randomly selected one hour from each of the 13 bins and computed the quotient of the fitted maximum to the true maximum for each hour. We then computed the three quartiles of the resulting quotients in each of the bins. Figures 7-10 show these three quartiles, plotted against the hourly average.

In contrast to the previous method, these estimators are nearly median unbiased. That is, the median value of the quotient is just about 1, corresponding to accurate estimation. For hourly averages greater than 1 ug/m^3 one can see that the iterated log models lead to estimates of the maxima

that are within 20 to 40% of the actual maxima at least half the time for the Kincaid data and within 10% at least half the time for the New York data. That is, the first and third quartiles of the fitted over actual ratios fall at .9 and 1.1 for New York, at .8 and 1.2 for Kincaid (at least on the right half of the plots). The log models have roughly the same error rates. It is also worth noting that, for the Kincaid data, the log models continue to give impossible fitted values in many cases.
Comparing these results to those obtained from the method of universal constants, one can see that the method of short-term monitors offers some improvement in accuracy over the former method, where the estimates are noticeably biased and errors of 20% in the estimated maximum occur half the time. The increased accuracy is much more noticeable with the New York data. At this time it is impossible to say whether a comparable difference in accuracy would be present at most population center stations and absent at most point source stations.

4.3 Errors in the Method of Time-Scale

The third method suggested was to remove a daily cycle from the observed hourly data and then assume that the relationship between peak and mean of twelve hourly readings is the same as the that in twelve 5-minute readings. A priori, one would expect that this method to be the least effective of the three. The correlation of successive 5-minute readings will be higher than that of successive hourly averages; averages over longer time scales should come from distributions closer to Gaussian so the functional form of the underlying distributions will not be the same. In fact, the parameter estimates obtained this way are seriously in error, as can be seen by comparing the estimates in Table 3 with those in Table 1.

Figure 11 shows plots of quotients of the maximum estimated from the 1-hour to 12-hour relation to the maximum estimated from the actual 5-minute to 1-hour relation. Results from both sets and both the log vs log and the iterated log vs log model are graphed. At high levels, the estimates in New York are too high by 10-20%; at low levels, they are seriously biased low. In the Kincaid data, estimates from the iterated log vs log model are too high by 50-60%; the performance of the log vs log model is even worse. These plots, which roughly correspond to the median accuracy using this method, were so bad that we did no further investigation for the Kincaid data.

A similar procedure was applied to the New York data to predict the maximum for the iterated log and log models, respectively, with results similar to those obtained from the Kincaid data. The predictions are biased high; three fourths of the time, the fitted value is at least 5 or 10% too high; half the time, the fitted value is at least 10 or 20% too high. Somewhat surprisingly, the log versus log model performs somewhat better than the iterated log versus log model for this data set.

5. Estimation of an Arbitrary 5-Minute SO2 Level

The second objective of the analysis was to find a model for the conditional distribution of an arbitrary 5-minute SO2 level, given the hourly SO2 average. As an alternative to the theoretical calculation, the following ad hoc method was considered.

1. Use deviations of 5-minute SO2 levels from their hourly averages, rather than the 5-minute levels themselves.

2. Make deviations from different hours comparable by dividing them by a suitable scaling factor. The usual scaling factors, the standard deviation or the interquartile range within an hour, cannot be used because one wants a method that can be used when knowledge of variability within an hour is not available. The scale factor must depend only on the hourly average. We employed a scale factor of the form exp(B *log(hourly average) + A). The slope and intercept, B and A, were obtained by OLS regression of log(hourly SD) on log(hourly average), in each data set separately. In practice, it would be necessary to use the parameter estimates from these two data sets in future data sets which contain only SO2 hourly averages.

3. Pool all the scaled deviations together and fit a simple parametric model to the resulting empirical distribution.

This three step method was applied separately to each data set. The estimated conditional distribution function is given by equation 6.

(6) Prob(5-minute SO2 level $\le x$; hourly average SO2 = y) =

$$F(x;\ y) = \Phi\left((x-y)/\exp(B*\ln(y) + A) \right).$$

The numerical values of A and B are given in table 4.

We found that the standard normal distribution worked acceptably well for both the New York and the Kincaid data. An attempt to use a three parameter gamma distribution to compensate for some skewness in the scaled deviations did not lead to enough improvement to justify the introduction of the extra parameters. One should note that there is a systematic error in this procedure that was not present in modelling of the maximum. Given the serial correlation of successive five-minute readings, the readings in the middle of the hour will be more highly correlated with the hourly average than will the first or last readings. The model in equation 6 is intended, at best, to predict the value of a 5-minute reading selected at random from one of the twelve time slots during an hour, not the value of a 5-minute reading from a specified time slot.

6. Error in the Estimation of Any 5-Minute SO2 Level

There are two types of error that one may consider here. First, there is the error in using equation 6 to estimate the proportion of 5-minute readings which exceed a given level of SO2. Second, there is the error in using the equation to estimate the level of SO2 that corresponds to a given percentile of the distribution of 5-minute readings. If one is concerned about the frequency of exceedances of a threshold for health effects, it is the first type of error that is of interest. We will discuss only the estimation of this first type of error.

Cross-validation between the two data sets was used to measure the error. The estimated slope and intercept of the scaling factor (the only unknown parameters in the model) from the New York data and the observed hourly averages from the Kincaid data to predict the scaling factors in the Kincaid data. We then divided all the observed deviations from the hourly averages by these scaling factors. If the parameter estimates are good, these scaled deviations should be close to a standard normal distribution.

We grouped these scaled deviations into 16 bins, according to the level of the hourly average. To quantify how well the estimates performed, we computed, for each of the 16 bins the observed proportion, β, of scaled deviations which exceeded the values -2, -1, -.5, +.5, +1, +2. This corresponds to using as thresholds the 5'th, 15'th, 30'th, 70'th, 85'th and 95'th percentiles of the 5-minute readings, computed using the correct parameters. Figure 14 shows the plots of these five β's against the hourly average. (The six curves correspond to the nominal 5'th through 95'th percentiles; the ordinate shows the percentage of scaled deviations actually less than that threshold.) The whole procedure was then repeated, reversing the roles of the New York and Kincaid data sets. Figure 15 shows the plots of the β's from New York data with Kincaid parameters.

It can be seen from these two plots that the 5-minute readings in the Kincaid data are more dispersed about their hourly averages than would be expected from the New York data. At high values of the average, a threshold which one would expect to be the 70'th percentile is actually only the 55'th to 60'th percentile; what one would expect to be the 85'th percentile is actually between the 60'th and the 70'th percentile; what one would expect to be the 95'th percentile is actually only about the 70'th to the 80'th percentile. Consequently, if one were using the New York data for parameter estimates, one would noticeably underestimate the frequency of exceedances of a threshold.

Necessarily, one finds the opposite situation when 5-minute readings in New York are inferred from the Kincaid data. As shown in figure 15, a threshold that one would expect, on the basis of the Kincaid data, to be only the 70'th percentile of 5-minute readings would actually be nearly the 95'th percentile in New York. Consequently, if one were using the Kincaid data for parameter estimates, one would noticeably overestimate the frequency of exceedances.

7. Theoretical Modelling of the Joint Distribution of 5-Minute Levels

We made some attempts to explore theoretically motivated parametric models for the third problem listed in the introduction, namely estimation of the joint distribution of the 5-minute levels, conditional on the hourly average. The most popular choice of marginal distribution for

SO2 levels, when averages over a single length of time are observed, is the lognormal. We therefore tested the goodness-of-fit of the lognormal distribution to the 5-minute sequences at Kincaid and New York. The 5-minute readings at New York appeared to fit a lognormal distribution acceptably. (A formal test would reject the hypothesis of lognormality. However, it appears that the deviation from the lognormal is small enough to be of no practical importance even though the enormous sample size leads to formal rejection of the model.) The 5-minute readings at Kincaid appeared noticeably more leptokurtic than a lognormal distribution. We therefore did no further work with the Kincaid data.

Estimation of the joint conditional distribution requires three further assumptions. First, we assume that the unconditional joint distribution of all the logs of 5-minute levels is multivariate normal. This seems reasonable in light of the approximate marginal lognormality. Second, we assume that the autocorrelation structure of the sequence of logarithms of the 5-minute levels is a simple serial correlation, the correlation at lag i being just rho to the i'th power. This is necessary to keep the number of parameters in the model down to three. In fact, the sample correlations at lags 2 to 4 are not too far from the second to fourth powers of the lag 1 correlation. Third, we assume that the hourly average observed was the geometric mean of the twelve 5-minute levels, although it was in fact the arithmetic mean. This assumption is explicitly false: the true geometric mean is smaller than the observed average, but the higher the correlation between successive 5-minute readings, the smaller the difference between the arithmetic and geometric means. This assumption is made in order to get an algebraically tractable problem and with the hope that the high serial correlation will make it close to true. With these three assumptions, it follows that the logs of the 5-minute levels and the log of their geometric mean come from a 13-dimensional normal distribution with a rank 12 covariance matrix.

One now finds that the desired conditional distribution of the vector of 12 log 5-minute readings, given the log of the geometric mean, is 12-dimensional normal with mean and variance given by the standard multivariate regression formulae. Letting Z_i = log of the i'th 5-minute reading, we have that the mean and variance-covariance matrix of this conditional distribution are given by equations 7A and B:

$$(7A)\quad E(Z_i \mid \bar{Z}) = \mu + \mathrm{Cov}(Z_i,\bar{Z}) * (\bar{Z} - \mu)/\mathrm{Var}(\bar{Z})$$

$$(7B)\quad \mathrm{Var}(\underset{\sim}{Z} \mid \bar{Z}) = \mathrm{Var}(\underset{\sim}{Z}) - \mathrm{Cov}(\underset{\sim}{Z},\bar{Z})\mathrm{Cov}(\underset{\sim}{Z},\bar{Z})'/\mathrm{Var}(\bar{Z}).$$

In more detail, the i'th coordinate of the vector of covariances of the logs of the 5-minute readings and the log of the geometric mean, $\mathrm{Cov}(Z_i,\bar{Z})$, is equal to

$$\sigma^2 * \{1 + \rho + \rho^2 + \ldots + \rho^{i-1} + \rho + \ldots + \rho^{12-i}\}/ 12$$

and the variance of the log of the geometric mean is equal to

$$V = \sigma^2 * \{ 12 + 2*[11\rho + 10\rho^2 + \ldots + \rho^{11}] \}/144.$$

The problem of estimating the joint distribution of the 5-minute levels, given the hourly average, is now reduced to the problem of estimating the three parameters (mu, sigma, and rho) in the above expressions, when one observes only the sequence of hourly averages. Because the sequence of observed logs of geometric means is also a multivariate normal sequence, it is simple to estimate the mean, variance, and covariance of this sequence. Specifically, the log of the geometric mean is normal with mean equal to mu, with variance equal to V above. Furthermore, the logs of the geometric means in successive hours are bivariate normal with covariance equal to

$$C = \sigma^2 * \{\rho + 2\rho^2 + \ldots + 12\rho^{12} + 11\rho^{13} + \ldots + 2\rho^{22} + \rho^{23} \} / 144.$$

The (computable) maximum likelihood estimates of the mean mu, variance V, and covariance C of the hourly averages uniquely determine the MLE's of the parameters mu, sigma, and rho of the 5-minute series.

The estimated conditional distribution of the logs of the 5-minute levels in New York, given their hourly averages, is shown in Table 5. This distribution is 12-dimensional normal with the indicated numerical values for the vector of conditional expectations of the logs of the 5-minute readings, given the hourly average, and for the variance-covariance matrix.

One can also attempt to elaborate on the above computation by making approximate corrections for the fact that one actually observes the arithmetic mean rather than the geometric mean. All of the above equations and distributional formulations are still valid. The only problem is that they cannot be used for computation if the geometric means are not observed. We suggest that the following approximations be used when only the arithmetic means are observed. First, compute the first and second sample moments of the observed sequence of arithmetic means and use these values to get method of moments estimates of the parameters mu, sigma, and rho. (The arithmetic means are not lognormal so these are not maximum likelihood estimates.) These parameter estimates then specify numerically the joint distribution of the 5-minute levels, given the geometric mean. To complete specification of this distribution, one need only give a numeric estimate, based on the arithmetic mean, of the geometric mean. A reasonable choice is to set the estimated sample geometric mean equal to the observed sample arithmetic mean times the ratio of the estimated expectation of the geometric mean to the estimated expectation of the arithmetic mean.

Application of the above protocol requires only expressions, in terms of mu, sigma, and rho, for four moments: the expectations of the sample arithmetic and geometric means, the variance of the sample arithmetic mean, and the covariance of the arithmetic means of successive hours. Given that the logs of the 5-minute readings are serially correlated normal(μ, σ^2)'s, the expected values of the arithmetic and geometric means are, respectively,

$E_A = \exp(\mu + \sigma^2/2)$ and

$E_G = \exp(\mu + \theta \sigma^2/2)$ where

$$\theta = \{ 12 + 2*[11\rho + 10\rho^2 + \ldots +$$

$$\rho^{11}] \}/144.$$

The variance of the arithmetic mean is

$$V_A =$$
$$\exp(2\mu + \sigma^2)*\{12 + 2*[11(\exp(\sigma^2\rho)-1) +$$

$$10(\exp(\sigma^2\rho^2)-1) +$$

$$\ldots + (\exp(\sigma^2\rho^{11})-1)] \}/144.$$

Finally, the covariance of the arithmetic means from two consecutive hours is equal to

$$C_A = \exp(2\mu + \sigma^2)*\{ (\exp(\sigma^2\rho)-1) +$$

$$\ldots + 11(\exp(\sigma^2\rho^{11})-1) +$$

$$12(\exp(\sigma^2\rho^{12})-1) +$$

$$11(\exp(\sigma^2\rho^{13})-1) +$$

$$10(\exp(\sigma^2\rho^{14})-1)$$

$$+ \ldots + (\exp(\sigma^2\rho^{23})-1) \}/144.$$

It is important to note that all of the above theoretical modelling is heavily dependent on the assumed multivariate lognormality of the 5-minute levels. If the 5-minute levels were marginally Weibull, Gompertz, or gamma then none of the above manipulations would work. Furthermore, in new data sets it will not be possible to check for lognormality of the 5-minute sequence by examining only the sequence of hourly averages. Thus, the techniques outlined in this section can only be applied by either taking lognormality on faith or by taking the trouble to observe enough 5-minute levels to perform at least a simple check on lognormality.

8. Conclusions

There does not seem to be any reliable method for estimating the maximum SO2 level within an hour from knowledge only of the time series of SO2 hourly averages at the same site. The theory that there is a simple relationship between the 5-minute and hourly averages, governed by the same constants at all sites, is not borne out by the two data sets examined. In fact, the functional form of the marginal distribution of 5-minute levels is not even the same at the two sites. One must recognize that the two sites considered were very different. The analysis should be repeated with data from similar sites to determine the extent of extrapolation across sites that is possible.

If the expense is not prohibitive, the best results are likely to be obtained by taking the trouble to measure the 5-minute time series for a period of 100 or so hours. Even this effort cannot promise better than an even chance of predicting future maxima to within ± 20%. Using parameter estimates from one of the few sites where 5-minute data have been collected or from the relationship between the hourly and 12-hourly averages at the site in question are likely to lead to somewhat less accurate predictions. The magnitude of the errors associated with attempts to predict the proportion of 5-minute readings which exceed a threshold are comparable to those experienced in estimating the maximum. If standards are to be established with the intention of limiting the health effects associated with high short-term exposures, then these limits on the accuracy in prediction must be borne in mind in the setting of standards.

Given the ad hoc nature of the parametric models used, one might try other parametrizations--e.g. estimate the transfer function between the time series of hourly averages and the time series of hourly maxima--to see if better approximations can be obtained. Because the iterated log model does a fairly good job of estimating the maxima in the data set from which the parameters were estimated and because the marginal distributions at the two sites considered are not even of the same form, we think it unlikely that other choices of parametrization will lead to much reduction in the cross-validation errors.

The task of estimating the conditional distribution of an arbitrary 5-minute level, given the hourly average, appears to be equally difficult. It appears that using ad hoc parameter estimates obtained from one site to predict 5-minute levels at another site leads to biased predictions. In the two data sets compared here, it was impossible to tell reliably whether a given level would be exceeded 5% or 30% of the time.

Estimation of the joint distribution of all twelve 5-minute levels, given their average, appears feasible only if one is prepared to assume a lognormal distribution for the unconditional distribution of these readings. There are data sets for which this is demonstrably not true. Thus, it again appears that the most reliable estimates can be obtained only by observing at least enough of the 5-minute sequence to check lognormality roughly.

BIBLIOGRAPHY

(1) Grandell, Jan (1984), Stochastic Models of Air Pollutant Concentration, Springer-Verlag, Berlin

(2) Johnson, Norman and Kotz, Samuel (1970), Continuous Univariate Distributions, vol. 1, John Wiley & Sons, New York

(3) Larsen, Ralph (1971), A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards, U.S. Environmental Protection Agency, Office of Air Programs, Research Triangle Park, North Carolina

(4) Legrand, Michael (1974), Statistical Studies of Urban Air Pollution---Sulfur Dioxide and Smoke, in Statistical and Mathematical Aspects of Pollution Problems, John Pratt ed., Marcel Dekker, Inc, New York

(5) Pollack, Richard I. (1975), Studies of Pollutant Concentration Frequency Distributions, U.S. Environmental Protection Agency, Office of Research and Development, Publication EPA-650/4-75-004, Research Triangle Park, North Carolina

TABLE 1

Descriptive Statistics

	Station	**Mean**	**S.D.**	**Skewness**	**Kurtosis**	**Maximum**
Hr Avg	**NY**	19.61	18	2.8	15	257
	Kincaid	20.78	75	47	3810	2500
Hr Sd	**NY**	3.34	3.6	3.5	21	57
	Kincaid	13.71	109	109	13000	5000
Log (Avg)	**NY**	2.64	.85	-.3	.3	5.55
	Kincaid	1.77	1.6	.0	-.2	7.82
Log (SD)	**NY**	.84	.84	.3	-.2	4.04
	Kincaid	1.25	1.4	1.01	.12	8.52

Regression of Log (Ratio) on Log (Average)

Station	**Slope**	**Intercept**	**RMSE**	**Ratio <1 When Average >**
NY	-.077	.499	.20	652
Kincaid	-.210	1.07	.69	163

Regression of LogLog (Ratio) on Log (Average)

Station	**Slope**	**Intercept**	**RMSE**	**Correlation**
NY	-.267	-.719	.62	-.34
Kincaid	-.258	-.191	1.06	-.36

TABLE 2

Distribution of Residuals at Kincaid

Value of Log(log(ratio))	**Percent**
-2.03	.05
-1.43	.10
-.70	.25
.23	.50
.76	.75
1.24	.90
1.43	.95

Table 3
Regressions from Method of Change of Time Scale

Model	Data Set	Slope	Intercept
Iterated	New York	-0.0854	-0.415
Log		-0.0528	0.716
Iterated	Kincaid	-0.12	0.606
Log		-0.170	2.010

TABLE 4

Fitted Models for Spread of 5-Minute Levels

Regression of Log (SD) on Log (AVG

Station	Slope	Intercept	Correlation Squared
NY	.687	-.972	.49
Kincaid	.645	.114	.53

Regression of SD on Average

Station	Slope	Intercept	Correlation Squared
NY	.114	1.109	.33
Kincaid	1.197	-11.169	.67

TABLE 5

CONDITIONAL MEANS AND VARIANCES OF LOG 5-MINUTE LEVELS

```
E(Z:zbar)     2.64   +0.984 *(zbar-2.64)
              2.64   +0.994 *(zbar-2.64)
              2.64   +1.002 *(zbar-2.64)
              2.64   +1.007 *(zbar-2.64)
              2.64   +1.012 *(zbar-2.64)
              2.64   +1.013 *(zbar-2.64)
              2.64   +1.013 *(zbar-2.64)
              2.64   +1.012 *(zbar-2.64)
              2.64   +1.007 *(zbar-2.64)
              2.64   +1.002 *(zbar-2.64)
              2.64   +0.994 *(zbar-2.64)
              2.64   +0.984 *(zbar-2.64)
Var(Z:zbar)
              0.056   0.041   0.027   0.015   0.003  -0.007  -0.015  -0.022  -0.027  -0.032  -0.034  -0.035
              0.041   0.043   0.028   0.016   0.004  -0.006  -0.014  -0.021  -0.026  -0.030  -0.033  -0.034
              0.027   0.028   0.032   0.019   0.007  -0.003  -0.011  -0.018  -0.023  -0.028  -0.030  -0.032
              0.015   0.016   0.019   0.025   0.012   0.002  -0.006  -0.013  -0.018  -0.023  -0.026  -0.027
              0.003   0.004   0.007   0.012   0.018   0.008  -0.001  -0.008  -0.013  -0.018  -0.021  -0.022
             -0.007  -0.006  -0.003   0.002   0.008   0.015   0.006  -0.001  -0.006  -0.011  -0.014  -0.015
             -0.015  -0.014  -0.011  -0.006  -0.001   0.006   0.015   0.008   0.002  -0.003  -0.006  -0.007
             -0.022  -0.021  -0.018  -0.013  -0.008  -0.001   0.008   0.018   0.012   0.007   0.004   0.003
             -0.027  -0.026  -0.023  -0.018  -0.013  -0.006   0.002   0.012   0.025   0.019   0.016   0.015
             -0.032  -0.030  -0.028  -0.023  -0.018  -0.011  -0.003   0.007   0.019   0.032   0.028   0.027
             -0.034  -0.033  -0.030  -0.026  -0.021  -0.014  -0.006   0.004   0.016   0.028   0.043   0.041
             -0.035  -0.034  -0.032  -0.027  -0.022  -0.015  -0.007   0.003   0.015   0.027   0.041   0.056

Here Z = vector of logs of 5-minute readings

zbar = observed value of log hourly geometric mean
```

FIGURE 1

MAXIMUM VERSUS HOURLY AVERAGE: NEW YORK DATA

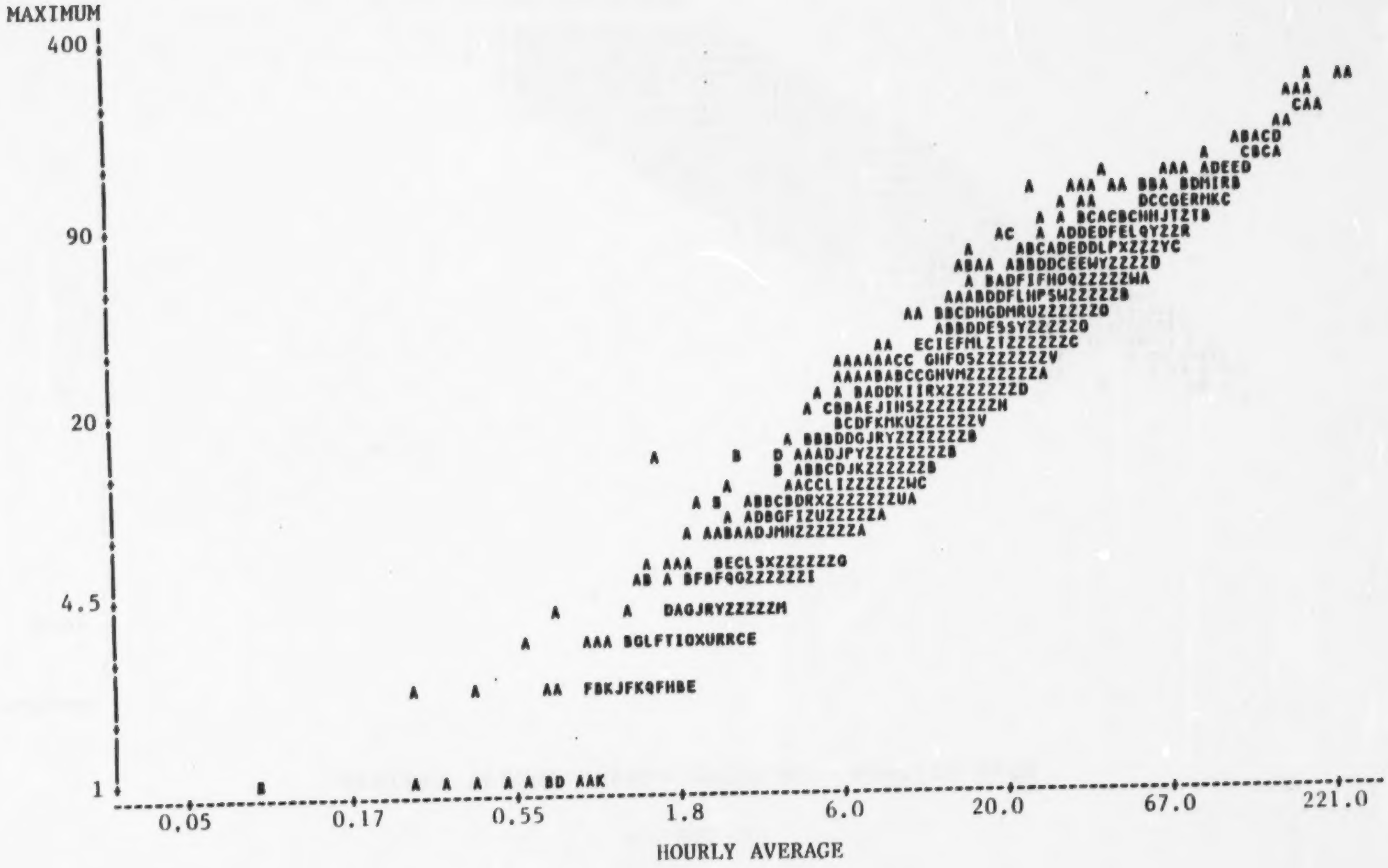


FIGURE 2

MAXIMUM VERSUS HOURLY AVERAGE: KINCAID DATA

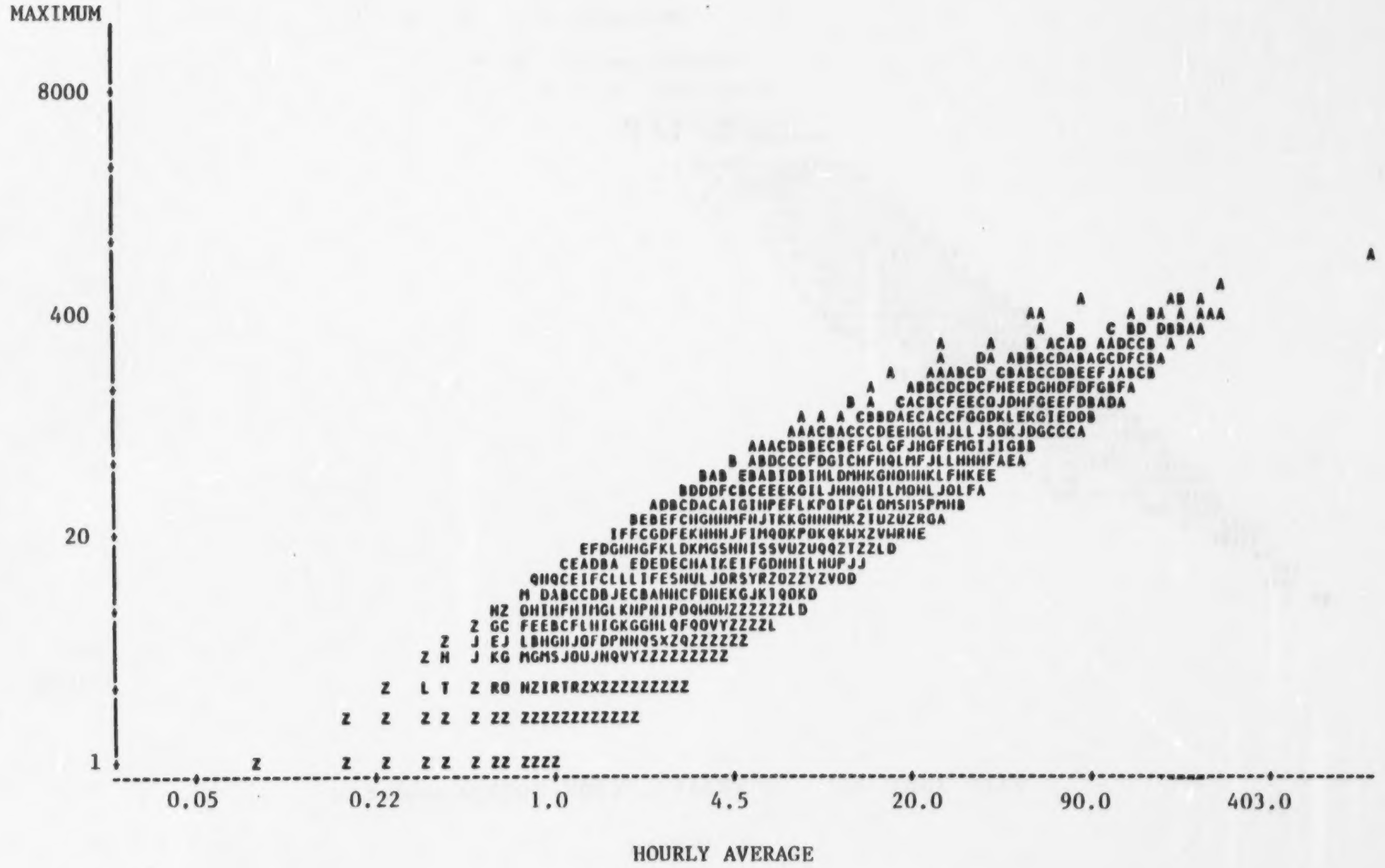


FIGURE 3

RESIDUALS OF ITERATED LOG MODEL: NEW YORK DATA

```
                                                       MISSING VALUE
                                                               COUNT
                                                        % COUNT/NOBS

       BAR CHART                                           # BOXPLOT
 2.25+*                                                    1
     .**                                                  91    0
     .*********                                          687    |
     .*************************                         2035    |
     .*************************************             3027 *--+--*
     .************************************************  3915 +-----+
     .************************                          1893    |
     .********                                           592    |
     .**                                                 124    0
     .*                                                   44    0
     .*                                                   13    0
-3.25+*                                                    4    *
      ----+----+----+----+----+----+----+----+----+---
      * MAY REPRESENT UP TO 82 COUNTS


    .
   36
 0.29

                    NORMAL PROBABILITY PLOT
 2.25+
     |                                                  *
     |                                         **********
     |                                 ********+
     |                         ********+
     |                **********
     |         *******+
     | +********
     |***
     |*
     |*
-3.25+*
      +----+----+----+----+----+----+----+----+----+----+
          -2        -1        +0        +1        +2
```

FIGURE 4

RESIDUALS OF ITERATED LOG MODEL: KINCAID DATA

```
                                                      MISSING VALUE
                                                              COUNT
                                                       % COUNT/NOBS

       BAR CHART                                           # BOXPLOT
 2.25+*                                                   14    |
     .**********                                         565    |
     .*******************************                   1868    |
     .************************************************  2891 +-----+
     .*******************************                   1852 *--+--*
     .******************************                    1782 |     |
-0.75+****************************                      1677 +-----+
     .********************                              1164    |
     .*******                                            380    |
     .*********                                          505    |
     .****                                               214    0
     .***                                                123    0
-3.75+*                                                    4    0
      ----+----+----+----+----+----+----+----+----+---
      * MAY REPRESENT UP TO 61 COUNTS

 2741
17.37
                     NORMAL PROBABILITY PLOT
 2.25+                                            ++++++*
     |                                        +++*********
     |                                  **********
     |                            ******+
     |                       *****++
     |                    ****++
-0.75+                ****+
     |           +*****
     |       ++****
     |  ++++****
     |++*****
     |***
-3.75+*
      +----+----+----+----+----+----+----+----+----+----+
          -2        -1        +0        +1        +2
```

FIGURE 5

ERRORS WITH USING FIXED ESTIMATES
(NEW YORK DATA, KINCAID PARAMETERS)

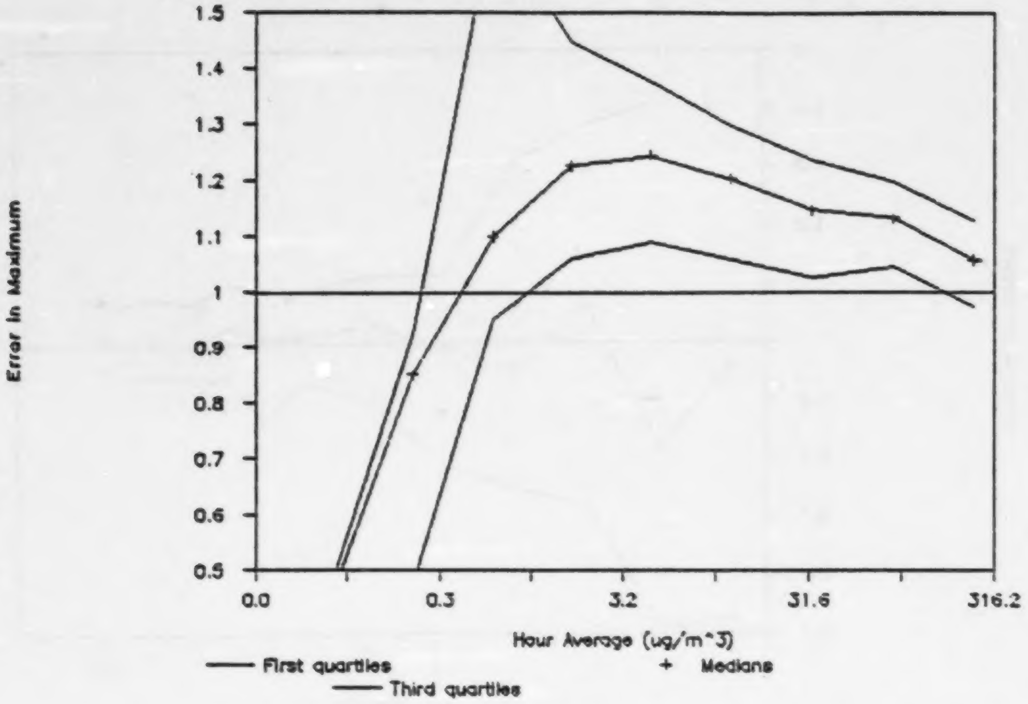


FIGURE 6

ERRORS WITH USING FIXED ESTIMATES
(KINCAID DATA, NEW YORK PARAMETERS)

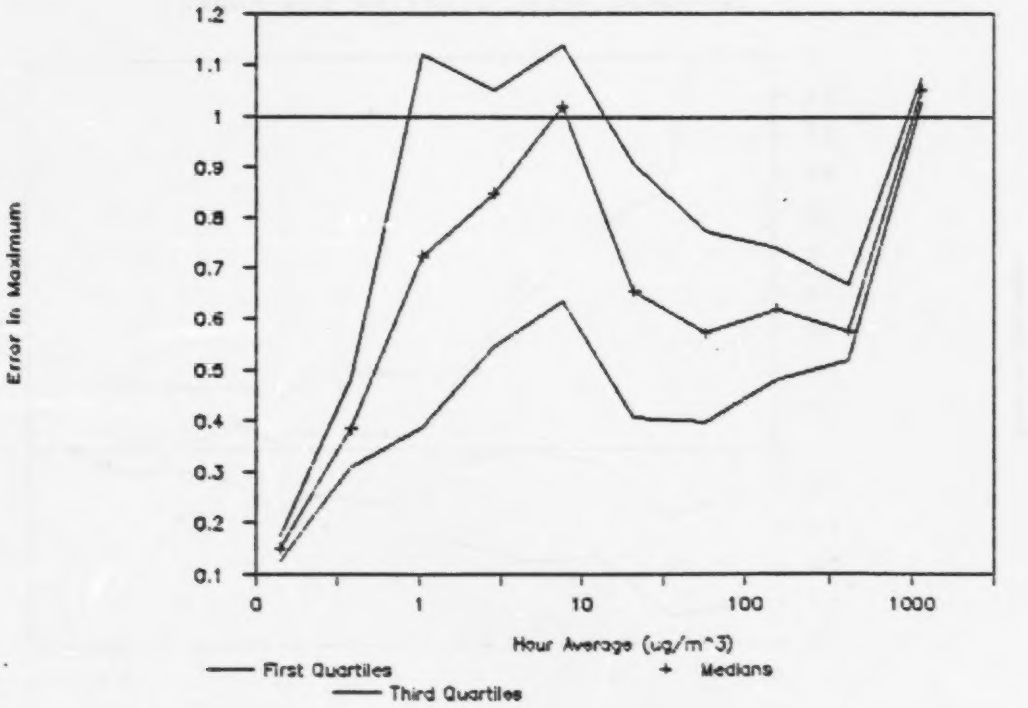


FIGURE 7

ERRORS WITH SHORT-TERM MONITORS
(NEW YORK DATA, ITERATED LOG MODEL)

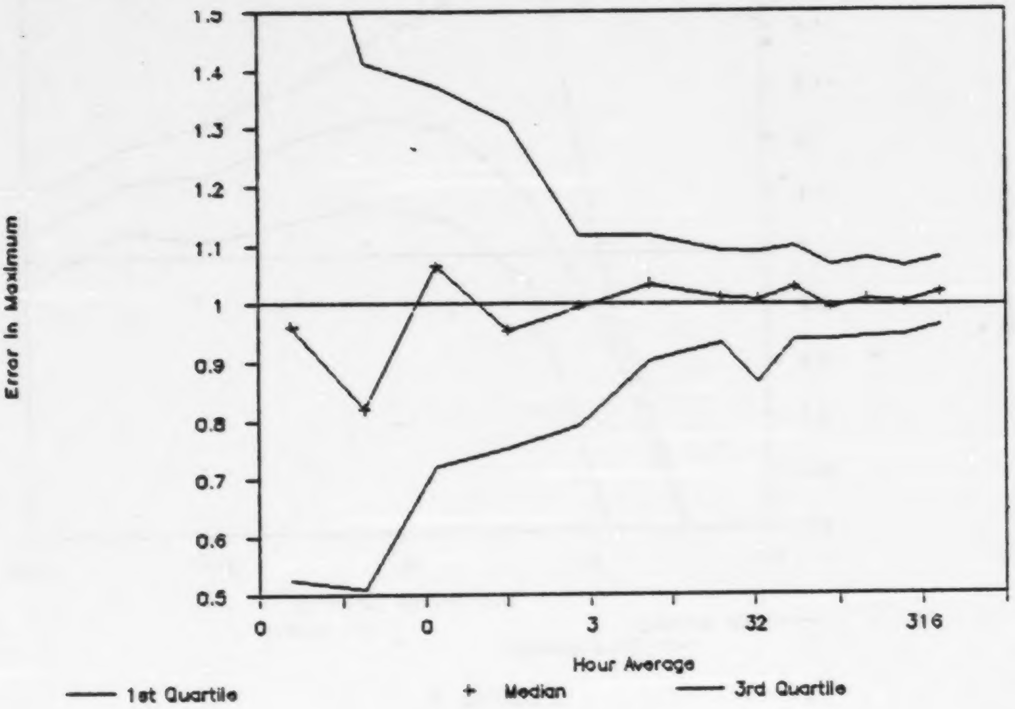


FIGURE 8

ERRORS WITH SHORT-TERM MONITORS
(KINCAID DATA, ITERATED LOG MODEL)

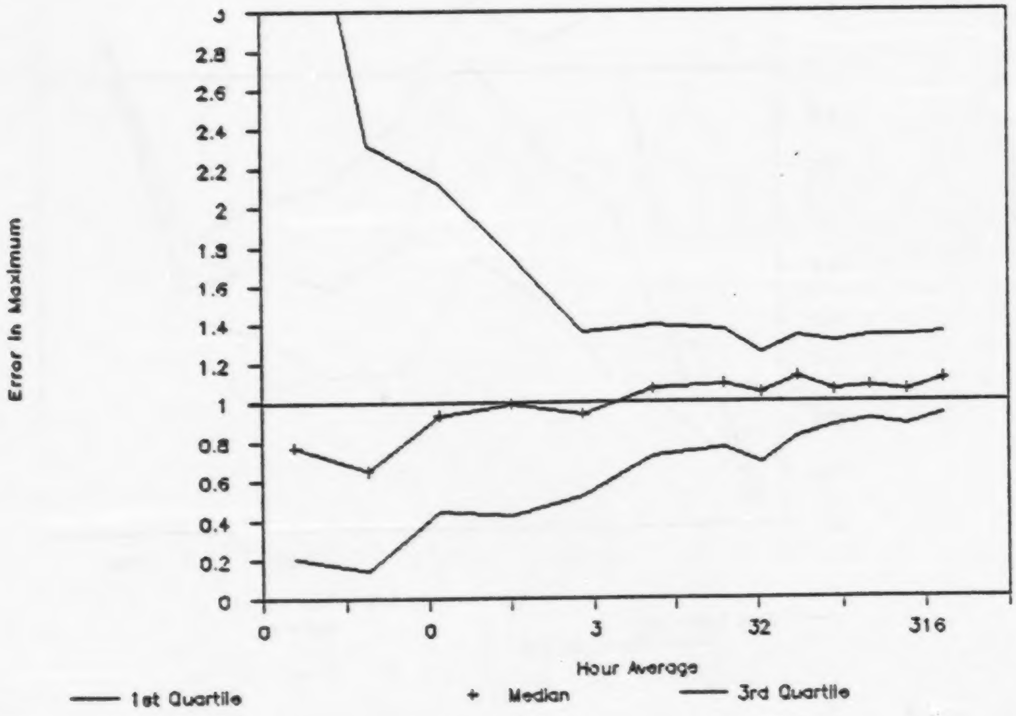


FIGURE 9

ERRORS WITH SHORT-TERM MONITORS
(NEW YORK DATA, LOG MODEL)

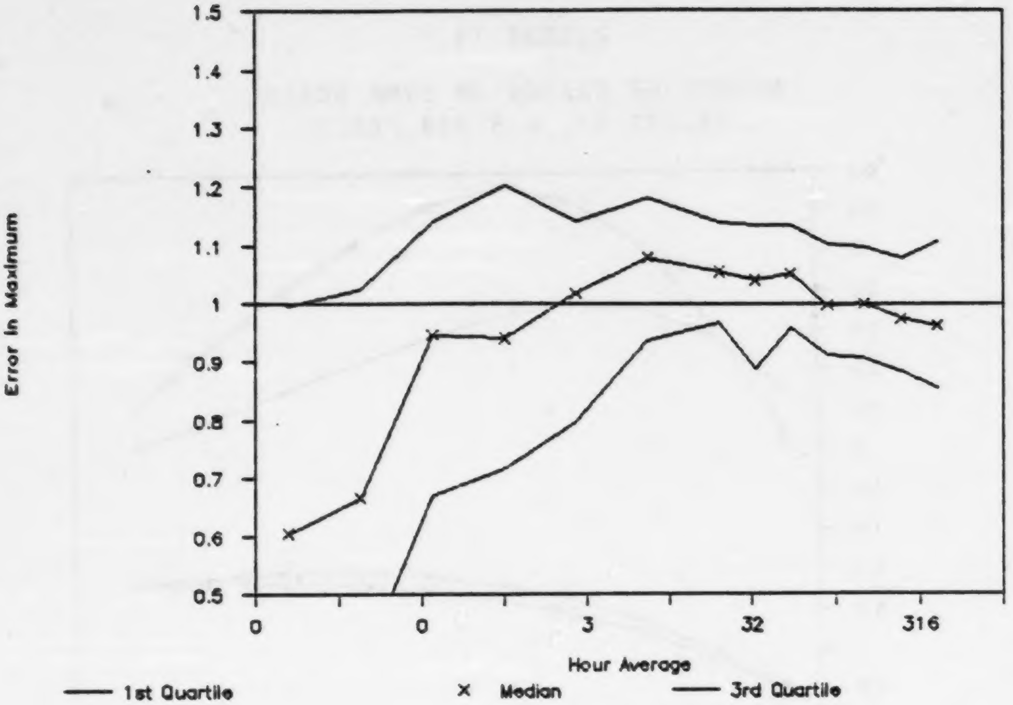


FIGURE 10

ERRORS WITH SHORT-TERM MONITORS
(KINCAID DATA, LOG MODEL)

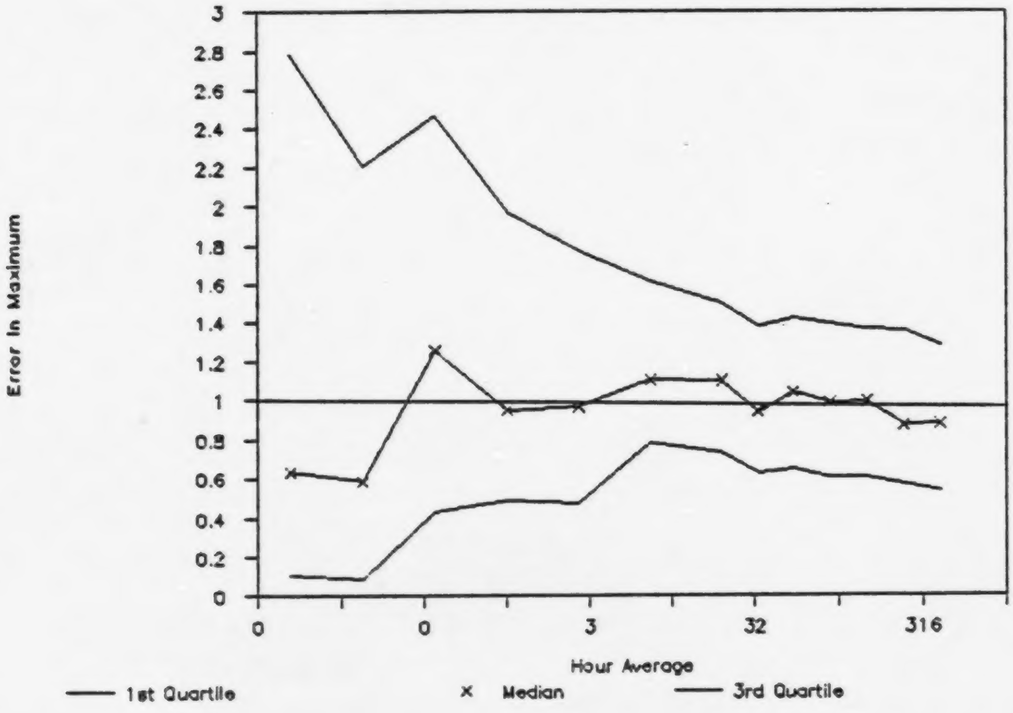


FIGURE 11

METHOD OF CHANGE OF TIME SCALE
(HR./12 HR. = 5 MIN./HR.)

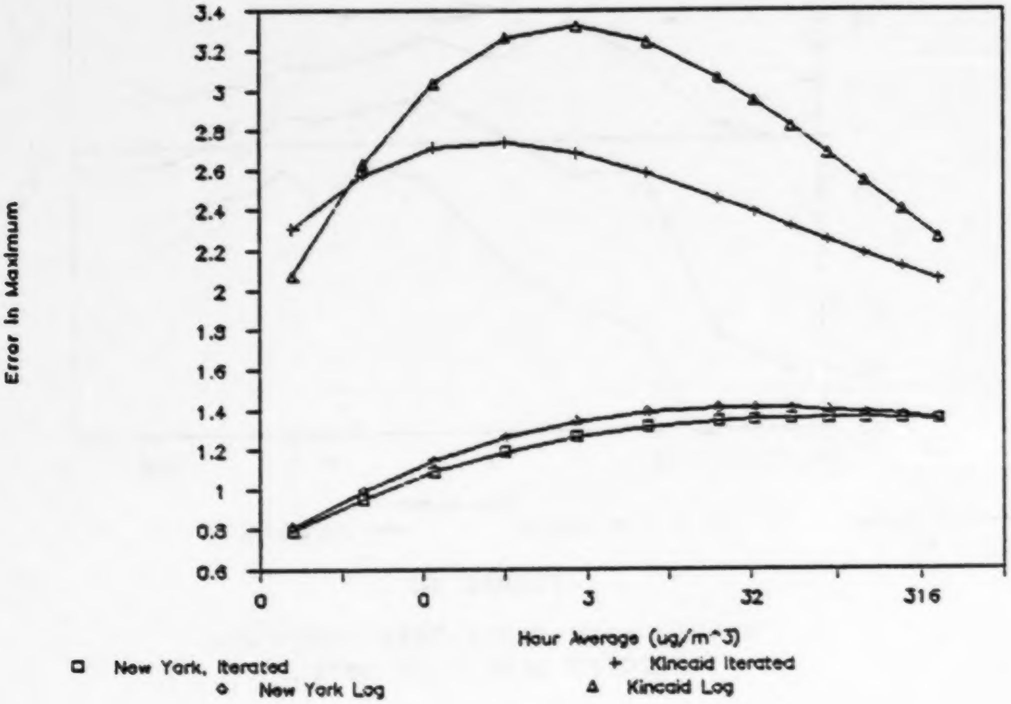


FIGURE 12

ERRORS WITH CHANGE OF TIME SCALE
(NEW YORK DATA, ITERATED LOG MODEL)

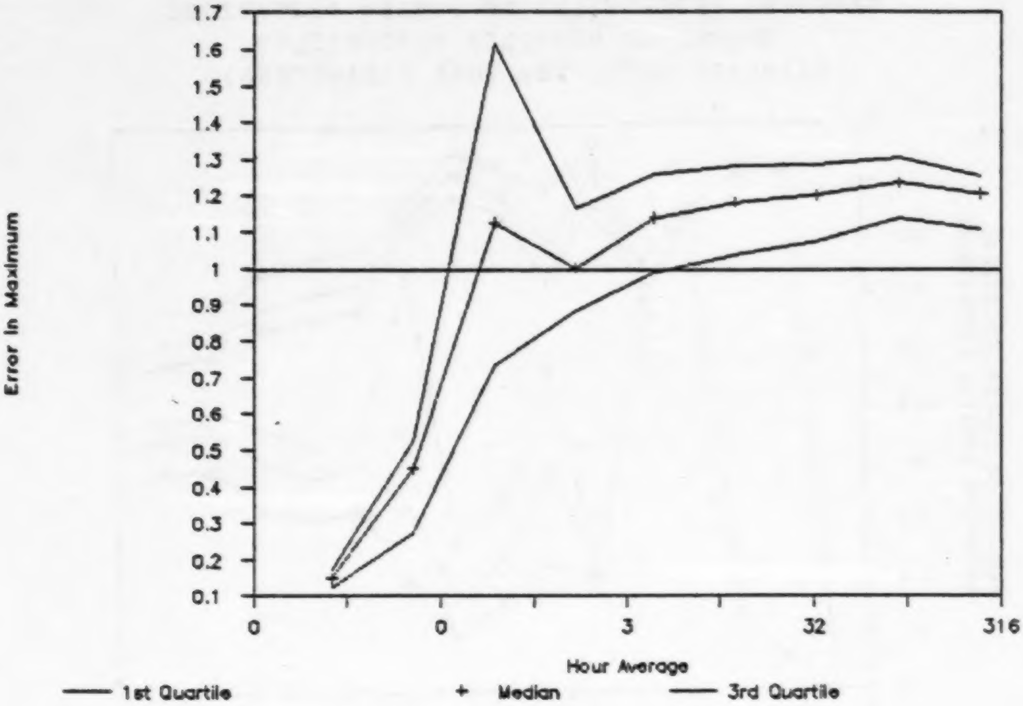


FIGURE 13

ERRORS WITH CHANGE OF TIME SCALE
(NEW YORK DATA, LOG MODEL)

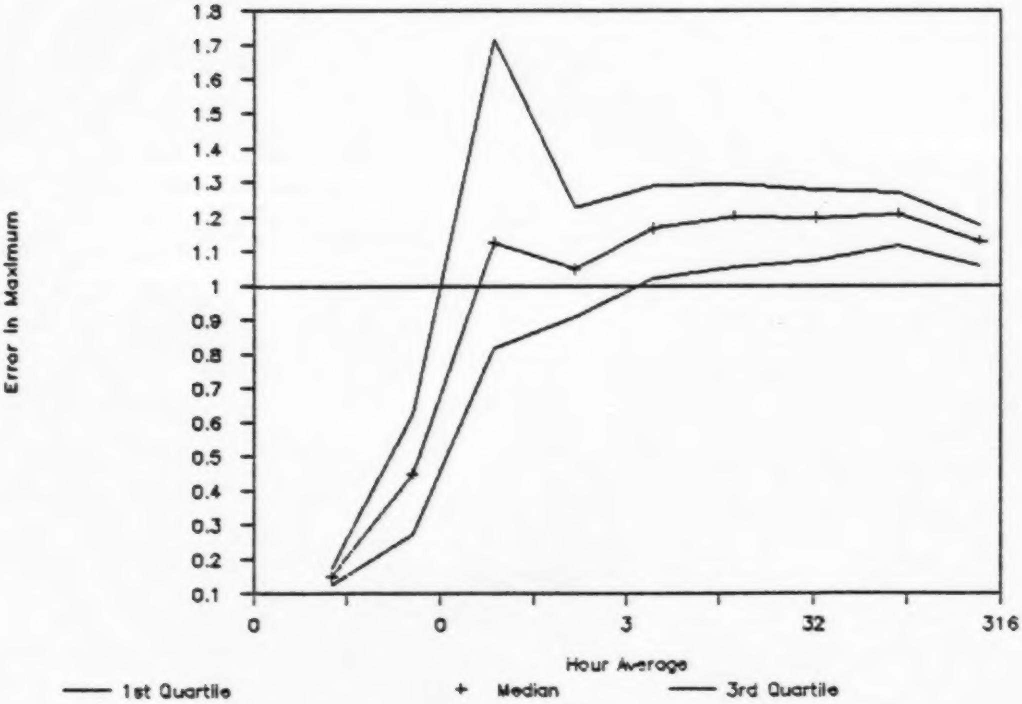


FIGURE 14

OBSERVED PERCENTILES OF SCALED DEVIATIONS:
GRAPHS OF MODELLED PERCENTILES
(KINCAID DATA, NEW YORK PARAMETERS)

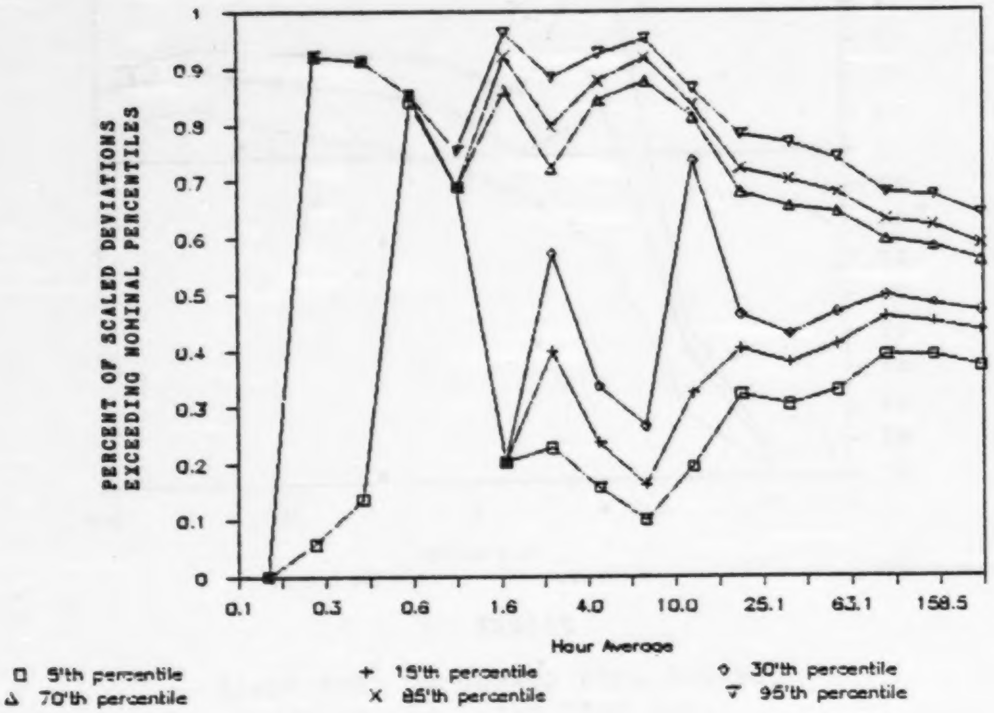
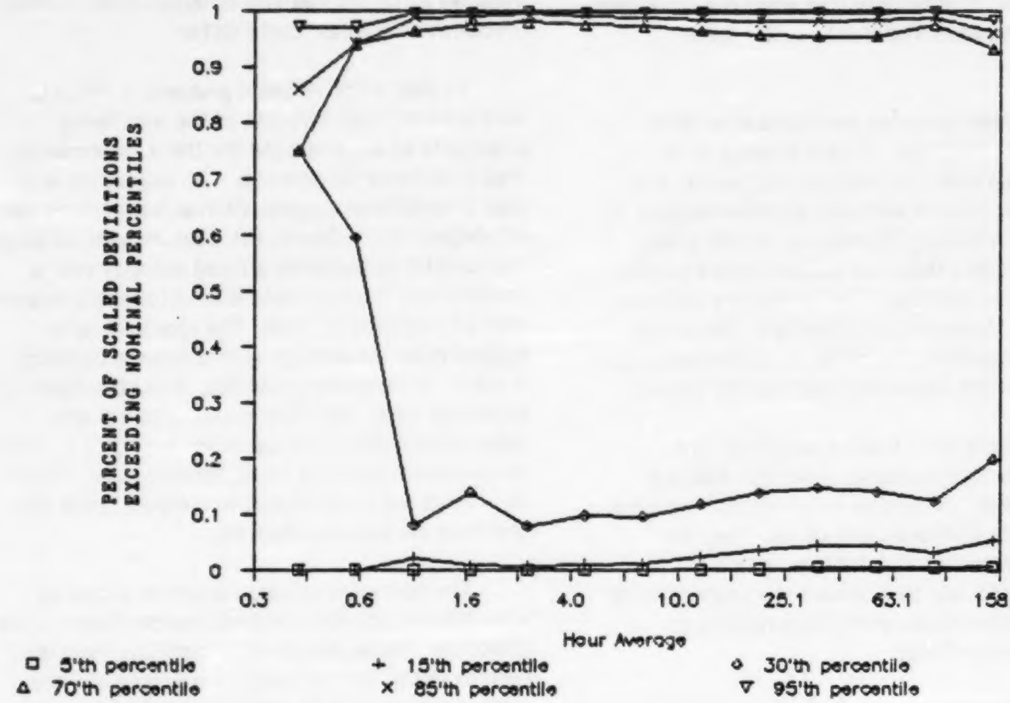


FIGURE 15

OBSERVED PERCENTILES OF SCALED DEVIATIONS:
GRAPHS OF MODELLED PERCENTILES
(NEW YORK DATA, KINCAID PARAMETERS)



DISCUSSION

R. CLIFTON BAILEY*
Health Care Financing Administration,
2-D-2 Meadows East, 6325 Security Blvd., Baltimore, MD 21207

A recent editorial suggested that there be no new data collection until present data sets are thoroughly analyzed. This a tough standard. Even if one attempted to thoroughly analyze present data sets there would always be the possibility for more analysis. This is especially true when one considers analyses based on multiple data sets - meta analyses.

The authors are to be commended for their extensive data analyses. Of course some of us remain disappointed that certain parametric and nonparametric models were not explored because of complexity. In stating the reasons for not doing certain analyses, I think the authors take a narrow view of what is possible. The issues may be more ones of cost, time or expected return. This in no way undermines the value of the extensive empirical exploration of the data undertaken by the authors.

The authors set a task of establishing a relationship between studies in which data are recorded in short, 5-minute, intervals and the more common choice of hourly summaries. They are especially interested in establishing this relationship because they believe it is necessary to have information on the short time records to establish health effects.

When the basic process is observed from several points of view--different measurements, such as the 5-minute and the hourly measurements, should be expressible in terms of the common process observed. The perspective of a common process being observed from different points of view provides the framework or model to work from From this perspective, distinctly different measurements or measurement processes generally are not equally informative of the process and the statistical properties of these measurement processes are not the same. In analyzing the data, it is important to remember that the measurement process is part of the observation and more than one quantity may be needed to describe the process. The model for the process generally will be a combination of stochastic and deterministic components. An issue underlying the effort to evaluate different methods of observation is that precision as well as costs differ.

To deal with the basic problem, it helps to have a model that consists of the underlying process to be observed and the the measurements used to observe the process. An evaluation with such a model may suggest alternative measurement strategies. For example, the measurement strategy may consist of obtaining a fixed quantity over a random time interval instead of obtaining a measure over a fixed time interval. The idea is clearly suggested by the analogy with a Poisson counting process. In counting statistics, two strategies are commonly used. One uses a fixed interval and obtains the count while the other specifies a count and measures the time to obtain this count. These strategies can be evaluated to compare costs and precision for a given situation.

The main concomitant measures explored were time of day and a meteorological factor, wind direction. These and other concomitant measures need to be part of the model. I would like to see more attention paid to concomitant factors at the two sites.

The authors state in their conclusions, " the theory that there is a simple relationship between the 5-minute and hourly averages, governed by the same constants for all sites, is not borne out by the two data sets examined."

The conclusions and recommendations are fundamentally sound. The authors recommend calibrating a model for each site. In this way differences among observed processes are properly recognized even if they are not explicitly modeled.

***Disclaimer**

The opinions are those of the author and do not necessarily reflect the opinions of the Health Care Financing Administration.

SUMMARY OF CONFERENCE

John C. Bailar III
Department of Epidemiology & Biostatistics, McGill University
Montreal, PO Canada H3A 1A2
and
Office of Disease Prevention & Health Promotion, U. S. Public Health Serv.
Switzer Building, Room 2132, 330 C Street, S.W., Washington, D.C. 20201

This summary of the conference is intended to provide some brief and integrated commentary on the eight papers and eight discussions presented here (1-16), plus some perspective on broader issues raised by the papers as a group but not covered by any one of them.

I will say much about unsolved problems. Of course, the more one knows about a situation, the easier it is to critique specific points and point to things that should be done. This is good for bringing out issues, but it can be bad if it creates an impression that problems dominate solutions. I do not want my comments here to be taken as a general indictment of compliance sampling, a field that has recently made much progress and is clearly making more.

Compliance Sampling in a Broader Context

The focus of the conference was compliance sampling; this term includes both a) the general assessment of how well we are doing in the management of hazards and b) the generation of data for individual action to enforce relevant laws and regulations. My basic view, as a citizen and scientist rather than a regulator, is that regulations should provide and should be interpreted as firm limits rather than targets, though they are often abused or misinterpreted as targets. Examples include the approaches of many states and cities to the control of criteria air pollutants, and the apparent attitude of parts of private industry that penalties for violations are a business expense, to be balanced against production volume and costs so as to maximize overall profits. Carol Jones (17) has commented on the effects of penalties on the probabilities of violations, and at this Conference Holley (11) has discussed such approaches in the context of bubbles.

But these two purposes of compliance sampling — overall assessment and enforcement — are broad and vague. There was very little said at the Conference about the ultimate purposes, or even the penultimate purposes, of these activities. This is a potentially serious gap, because what we do (or should do) in compliance sampling can be profoundly affected by matters beyond the short term goals of accurate assessment of the distribution and level of specific hazardous agents. Is our ultimate goal to protect human health? If so, what does that mean for the design of a program in compliance sampling, given our limits on time, money, attention, and other resources? How are concerns about cancer to be balanced against concerns about (say) birth defects, or heart disease? How are concerns about health in the U.S. to be balanced against health in other countries? How are we to balance short-term protection of our own health against protection far into the future, even across generations not yet born? How should we view and assess the quality of outdoor (ambient) air vs. indoor air (Hunt, 4)? There are similar very broad questions about direct health effects vs. the indirect health effects of unemployment and poverty, or restricted choices of important consumer goods, on protection of health. How are such matters to be developed in a context of concern about protection of non-health values, such as limiting the role of government in controlling private behavior or in facilitating compensation for harm actually inflicted (perhaps at much lower overall cost to society), the effects of unenforced or unenforceable directives on respect for the law in other areas, and many other matters? I recognize that such issues are generally to be dealt with at the highest political and social levels, but their resolution can have a profound effect on compliance sampling, and compliance samplers should understand the issues and express themselves as knowledgable professionals. Whether an inspector chooses to return to a plant that was in violation last month or to visit a new plant may depend on how much the agency depends on quiet negotiation vs. threats of legal action. Whether limited resources are used to sample for agents with acute, lethal, and readily identifiable toxicity or for more common but less characteristic and less devastating chronic disease may depend on what recourse is available when injury is suspected. Intensity of sampling (and of enforcement) in some critical industry may even depend on the state of the industry, and the state of the economy more generally.

The importance of defining the goals of compliance sampling in the broadest way is clear. But we have not dealt very well even with defining goals at more technical levels. Suppose that a well-conceived regulation sets a maximum exposure limit of 10 ppm. Should compliance sampling be designed to give only a yes/no answer, perhaps expressed as a Bernoulli variable, about whether some stream, or factory, or city is in violation? Should we instead try to determine the mean exposure over some defined region of time and space? The mean and variance, or the tails generally? Should we go only for the order statistics, especially the extremes (which will generally provide a moving target as problems are solved and compliance improves)? Do we need the whole probability distribution of values? Surely a yes/no answer can lead to much nonsense, as it did in some erroneous interpretations by the news media of a recent NAS report on drinking water, and some aspects of the probability distribution of values need more attention than others, but surely there is also a point where we have learned enough about that distribution, and must invest additional resources in the study of other problems.

Gilbert et al. illustrate this general need for precise goals in their discussion (9) of sampling soil for radioactivity. Was the underlying goal to determine whether radiation levels at any square inch of surface were above the standard? Was it to average, or integrate, over some unspecified larger area? Was it to determine means and variances, or other aspects of the distribution? Here, maybe the goal was in fact to determine means for small areas, but we would still need to know more about the problem, especially about the small-scale variability of contamination, to determine an appropriate sampling plan. For example, if contamination is nearly uniform within each area for which a mean is required, one test per sampled area may be enough. Conversely, if there is much chance of having one very small, very hot rock (of, say, 10^3, 10^4, or 10^5 pCi/g) one might have to sample on a much finer grid. The general issue here is the scope or range for averaging (or otherwise "smoothing") results. Chesson (10) has also commented on needs for relating statistical procedures to specific problems and contexts. Holley's work on the bubble (11) deals with a kind of averaging, but this Conference as a whole has given rather little attention to even this level of goals.

Likewise, there was little discussion of how strategies for compliance sampling must accommodate the likelihood of legal challenge. A probable freedom from such challenge may well have given Gilbert (9) considerable latitude to be complex, to use a great deal of peripheral information, and to interpret EPA's raw standard as he settled on the scope and distribution of sampled areas, to decide that he could ignore possible variation over time, and to develop a special sampling protocol.

At this point, one may begin to wonder about the role of statistics (and statisticians) in compliance sampling. I believe very strongly that the most visible, and apparently the most characteristic aspects of statistics – modeling of random variation, algebra, and computation – are only a small (though essential) part of the field. Statistics is, rather, the art and science of interpreting quantitative data that are subject to error, and indeed, in the study of environmental hazards, random error may account for only a tiny part of the uncertainty. Ross discussion (12) brings out clearly the real potential of statistics in the design of bubbles as well as the way bubbles ignore some important distributional issues.

I turn now to three sets of generic problems in compliance sampling: those in policy and concept, in unpredictable (stochastic) influences on the data, and in applications of theory. These sets of problems are broad and deep, and statistical thinking has a large and critical role in each.

Policy and Conceptual Aspects of Compliance Sampling

The first set is related to policy and concepts. I have already referred to the differences between broad public goals and more narrowly statistical goals, but there are many intermediate questions about what it is that one wants to accomplish, and what is feasible.

Approaches to evaluation in many fields fall rather well into three categories: evaluation of structure, of process, and of outcome. Each can be defined at multiple levels, but here it may be most useful to equate structure to the chemical methods, engineering and mechanical structures, and other aspects of the generation of hazardous agent; process to the emission or other release of the hazard into the community, its transport after release, and exposure levels where people are in fact exposed; and outcome to the human health endpoints (or other endpoints) that are the more fundamental objects of concern. Compliance sampling focuses on process (in this context), but it is not clear that there has been much hard policy thinking about whether this is the best way to attain the still rather fuzzy goals of the activity.

One aspect of this matter is the need to consider sensitive subgroups of the population. Such subgroups may not always be evident (as seems likely with some carcinogens), and their existence may not even be suspected, but somehow we must recognize not only that some people get sick from exposures that do not affect others, but that not all persons have the same probability of responding to some toxic agent.

A related point is "conservatism" in regulation, and its reflections in compliance sampling. Conservatism has several purposes, including the protection of sensitive subgroups, and the need to provide a cushion against random and nonrandom excursions of exposure to higher levels. I believe that its main use, however, is to protect us against our ignorance, not against our failures. We simply don't know what goes on within the human body at low exposure levels of carcinogens and other toxic agents, and choice of the wrong statistical model could lead to risk estimates that are wrong by orders of magnitude. Unfortunately, underestimates of risk will tend to be far more serious than overestimates if one works on a log scale, as is implied by the phrase "orders of magnitude." Implications of conservatism for compliance sampling are substantial. It does little good to set conservative limits for exposure if sampling, and hence enforcement, do not follow. It is not at all clear that regulatory agencies have been consistently attentive to the logical link between conservatism in risk assessment and conservatism in enforcement; indeed, some agencies may have it backwards, and believe that conservative exposure limits actually reduce the need for compliance sampling. There is scope here for a new study of how to trade off the risks and costs of (say) a higher exposure limit plus more rigorous sampling to assure compliance vs. a lower exposure limit that is to be less vigorously enforced.

Another policy and conceptual issue in compliance sampling has to do with distributional effects. When dose-response curves are linear at low doses, the mean exposure level in a population determines the expected number of adverse events, but it may still matter a great deal how the risk is distributed over the population. For example, it is no longer acceptable (at least in the U.S.) to concentrate the risks of toxic exposures on the lowest economic and social groups. Nor does one often hear arguments in favor of placing a new toxic hazard in an area already contaminated on grounds that a little more would not make much difference, even though this might be rational if there is reason to think that the risk is concentrated on a small, sensitive subpopulation that has already been "exhausted" by prior exposures.

Time does not permit more than a listing of some other policy issues in compliance sampling. How should ambient "natural" exposures to some agents, such as ozone, be accommodated in protocols for compliance sampling? What do we mean, in operational terms, by an "instantaneous" exposure? Marcus gave a strong start to the conference with his discussion of the need to design compliance sampling programs in light of the different time scales for environmental exposure, biologic response, and regulatory action (1), while Hertzberg (2) has pointed to some of the practical problems of doing so. How should, or how can, model uncertainty be built into sampling plans, including models of distribution and exposure as well as models of outcome?

Stochastic Aspects of Compliance Sampling

Issues to this point have not depended on any aspect of uncertainty in measurement or on random variability in the substance understudy. The steps from a precise deterministic model to an uncertain stochastic model introduce new issues. What are the roles of deterministic vs. stochastic models, and how should those roles affect compliance sampling? It is perhaps understandable that in enforcement actions, compliance data are treated as free of random variation, but surely this matter needs some careful thought.

Another issue arises from gaps in the data — gaps that are sometimes by design and sometimes not. There was little attention to this matter in this conference. Though every applied statistician is familiar with the problem, fewer are aware of the theoretical and applied approaches that have been worked out in recent years. These range from modeling the whole data set and using iterative maximum likelihood methods to estimate missing values (the E-M algorithm) to the straightforward duplication of some nearby value, which may be in error but not as far off as ignoring the missing observations, which in practice generally treats them as if they had the mean value for that variable ("hot deck" methods). Little and Rubin (18) provide an introduction to this topic, and techniques analogous to kriging, a method often used in geostatistics, may also be useful (19).

Unfortunately, the probability distributions of greatest interest in compliance sampling may often be hard to work with at a practical level. They tend to be "lumpy" in both space and time, with extreme variability, long tails to the right, and big coefficients of variation. Correlation functions over space and time (as in kriging) are important, but may themselves need to be estimated anew in each specific application, with detailed attention to local circumstances.

One practical consequence of dealing with "difficult" distributions is the loss of applicability of the Gaussian distribution (or at least loss of some confidence in its applicability), even in the form of the central limit theorem. Another is the loss of applicability of linear approaches, which have many well-known practical advantages with both continuous data and discrete (even non-ordered) classifications. Nonlinear analogs of, say, the general linear model and the loglinear or logit approaches have neither the theoretical underpinnings, nor the range of packaged general-use computer programs, nor the background of use and the familiarity of the linear approaches.

Given a set of data and a need to "average," what kind of average is appropriate? Some obvious questions have to do with ordinary weighted averages, others with moving averages. Still other questions have to do with the form of the averaging function: arithmetic, harmonic, geometric, etc. Geometric means are sometimes used in compliance sampling, as Wyzga has noted here (15), but they may often be quite unsuitable precisely because their advantage in some other situations - that they reduce the importance of high outliers - obscures the values of most concern. When health is at issue, I want a mean that will attend more to the upper tail than the lower tail. If six values on six successive days are (for example) 1, 2, 3, 4, 6, and 12, the geometric mean is 3.46, distinctly less than the arithmetic mean of 4.67, but it is the 6 and 12 that may matter most. An average that works opposite to the geometric mean seems better, such as the root mean square (5.92 in the example above) or root mean cube (6.99 above). I was glad indeed to learn recently that the geometric mean has been abandoned in measures of air particulates.

Many statistical approaches incorporate an assumption that the variance of an observation is independent of its true value. This may rarely be the case. However, lack of uniformity in variance may often have little consequence, and in some other cases it can be readily dealt with (such as by log or square root transforms). But there may be serious consequences if the nonuniformity or the statistical methods have statistical properties that are not understood, or are not acceptable. For example, in the 6-value numeric example above, if variances are proportional to the observed values, a log transform may produce values of approximately equal variance; however, the arithmetic mean of logged values is equivalent to the geometric mean of the original values, so that a different approach may be better. Problems are even greater, of course, when it is biases rather than random error that may depend on the unknown true values. Nelson's paper here (3) is rich in these and other statistical questions as well as policy questions.

Empirical Aspects of Compliance Sampling

The compliance sampler must attend to a wide variety of issues of direct, practical significance that derive from the context in which the data are to be collected and used. One is that results must be prepared so as to withstand legal challenge and, sometimes, political attack. A practical consequence is that much flexibility and much scope of application of informed judgment are lost. There may also be extra costs for sample identification, replicate measurement, and extra record keeping that help to validate individual values but reduce resources for other sampling that may contribute as much to the public health. This is in part a consequence of competing objectives within the general scope of compliance sampling. What is the optimum mix of finding indicators of many preventable problems and applying gentle persuasion to remove them vs. nailing down a smaller number of problems and ensuring that the data can be used in strong legal action if need be?

A second broadly empirical issue is the whole range of chemical and physical limitations on the detection and accurate measurement of hazardous substances. This is not the problem it once was — indeed, some observers believe that increased sensitivity of methods has led to the opposite problem of overdetection and overcontrol — but some substances are still difficult to measure at low concentrations by methods that are accurate, fast, and inexpensive. Thus, measurement remains a serious problem. An example is USDA's program for assessing pesticide residues in meat and meat products, which is limited by high costs to about 300 samples per year for the general surveillance of each major category (e.g., "beef cattle.") Thus there is a close link between the setting of standards (what is likely to be harmful, to whom, in what degree, and with what probability?) and the enforcing of standards (what violations are to be found, to what degree of precision, and with what probability?). A standard not enforceable because of limits on laboratory methods is no better, and may be worse, than no standard at all, and should be a candidate for replacement by some other method of controlling risk (e.g., process standards, or engineering controls). Sometimes, of course, deliberately insensitive methods can be cultivated and put to use. An example is FDA's "sensitivity of the method" approach to carcinogens in foods. Another real example, though slightly less serious here, was the step taken by the State of Maryland to improve its performance in enforcing federal highway speed limits: Move radar detectors from the flat straightaways to places where many drivers slow down anyway, such as sharp curves and the tops of hills, as other states had done long before. The incidence of detection of speed violations dropped markedly, and Maryland was suddenly in compliance with Federal standards. Creative design of a compliance sampling plan can produce pretty much whatever the designer wants, and I take it that a part of our task here is to develop approaches that discourage, inhibit, and/or expose the cynical manipulation of sampling procedures.

Sometimes, methods exist but for other reasons the data have not been collected. One example is the distribution of various foreign substances in human tissues. These include heavy metals, pesticides, and radioactive decay products; none of these had been adequately studied to determine the probability distribution of body burdens in the general population. Reasons are varied and deep, but include cost, problems of storage, control of access to banks of human tissues (an expendable resource), and ultimately the problems of procuring enough of the right kind of material from a fully representative sample of people. The need for detailed human data will surely grow with the growth of new approaches to risk assessment (especially of carcinogens), and compliance sampling may well be involved. Toxicokinetics, in particular, often demands human data; mechanisms can be examined in other species, but human sensitivity, human rate parameters, and human exposure can be determined only by study of human circumstances and, sometimes, human specimens.

Compliance sampling is indeed an activity loaded with problems. Overall, there is a clear need for substantially more thought and research on the empirical issues raised by compliance sampling. Wyzga (15) and Bailey (16) provides a fresh view of many of these.

Overview of the Overview

Where do we go from here? It is easy to call for more and better compliance sampling, and to show how we could then do more and better things. That will not get us far in this age of constrained resources. I believe that we need some other things first, or instead.

First is a broader and deeper view of compliance sampling. Many agencies and programs do such sampling, but almost always with a narrow focus on the enforcement of one or another regulation. This view should be broader — to include other substances, other agencies, and other objectives (including research) — and it should be deeper, so that issues of compliance sampling are considered at each stage from initial legislation onward, and plans are integrated with all other relevant aspects of Agency activities. Compliance sampling simply must not be treated like a poor relative — tolerated but not really welcome, and largely ignored until its general shabbiness or some genuine scandal forces a response.

A broader view of compliance sampling might, for example, support Nelson's comments on extensions from existing data to broader groups, even to national populations (3). Nelson's paper as a whole is unusually rich in both statistical questions and policy questions. While the matter seems to have received little specific discussion, it seems to me that the maximum useful geographic range or population size for compliance sampling, and maybe the optimum too, is the same as the maximum feasible scope of specific control measures. Thus, national data may be most critical in drafting or revising national laws and regulations, but local data are indispensable for understanding local needs, monitoring local successes, and enforcing local sanctions.

Another aspect of broadening our view of compliance sampling is the need to optimize sampling strategies for attaining specific, carefully elaborated goals. Thus, there might be reason in public policy to extend the use of weighted sampling, with more effort to collect samples likely to be out of compliance. This approach seems to have substantial informal use, especially when inspectors have considerable latitude to make decisions in the field, but has had less in the way of formal attention.

Still another aspect is the need for empirical study of the probability distributions that arise in the samples, and the development of sampling plans and analytic approaches that accommodate those distributions. Should one take a "point" sample of just the size needed for testing, or take a more distributed sample, mix it, and test an aliquot? Is there a larger role for two-stage sampling, in which the selection of a general area for examination is followed by the selection of sub- areas? Or a role for two-stage testing, in which aliquots of several samples are mixed and tested for the presence of some offending substance, with further testing of individual samples only if the group result is positive?

Perhaps the most fundamental need in developing a more comprehensive view of compliance sampling is for careful consideration of the role of genuinely random sampling, as opposed to haphazard or subjectively selected samples of convenience. One of the biggest surprises to me at this Conference was the lack of attention to the need to guarantee genuinely random sampling, though it provides the only acceptable justification for the statistical measures, such as p-values and confidence limits, that have been tossed about quite freely here. As a part of this, there is a clear need for new approaches to the computation of variances and other functions of the data, which will force demands for some kinds of randomization in the sampling. Gilbert's problem in particular (9) calls for highly sophisticated statistical modeling and analysis.

Second is a deeper consideration of how compliance sampling can be made more productive than in just the detection of violations, and how it can support broader Agency and national objectives. I have already referred to several aspects of this, but some points still require comment. One is the value of designing compliance programs (including sampling) that encourage both more and better monitoring and also encourage what might be called supercompliance. Response to the findings of a particular sample or pattern of samples may be yes-or-no, but surely one should put greater weight on finding the bigger violations. Frank has referred to this (13), with special comment about the potential value of variable frequency (and intensity) in sampling, while Warren (14) has noted some practical obstacles.

Some statistical tools do exist to aid in increasing the broad utility of data from compliance sampling. Bisgaard (5) and Price (7) have each presented reasons for more careful attention to the operating characteristics (OCs) of programs for compliance sampling. OCs might in fact be a good way to communicate with Agency administrators and others about the consequences of choosing one or another approach to monitoring, though Johnson (6) has emphasized the need for attention to the upper tail of exposure rather than the mean. It seems to me that the question of tail vs. mean may well depend on the health endpoint in question; an effect such as cancer that is considered a function of lifetime exposure may well be approached by means, while effects that really depend on short-term peaks should be regulated in terms of peaks, though this may create some problems when both kinds of endpoints must be managed in the same exposure setting. Bisgaard and Hunter (5) are firmly on the right track with their insistence on a more comprehensive view that integrates sampling protocols, calibration of the tools and processes, and a decision function to determine responses. This also underlines the need for clear articulation of goals; otherwise, Bisgaard's approach cannot be implemented. Johnson (6) also points to the need for adequate attention to other matters, too, including the political situation, pollutant behavior, sampling constraints, and the objectives of the standard. Flatman (8) also emphasizes the need for constant attention to the practicalities of solutions to real, and different, problems.

Other statistical tools of potential value in compliance sampling can be found in the epidemiologist's approach to diagnostic testing, with an insistence that policy decisions about testing be based on sound data on sensitivity, specificity, and positive and negative predictive values. These concepts have proved invaluable in policy decisions about medical screening, and they have similar potential to sharpen decisions about environmental screening.

Third, and my final point, is a plea that regulatory agencies explore the potential of statistical decision theory in their approaches to compliance sampling, including explicit consideration of the value of new information. The emphasis this will put on such matters as prior distributions, objective functions, cost functions, and balancing of disparate endpoints — all of which are already major elements in setting policy about compliance sampling — can only be good. Among other benefits, decision theory will tend to direct Agency attention to those points where the biggest improvements can be made, and away from both fine-tuning of little things with little potential profit and spinning wheels over big things that can't be settled anyway.

This would again direct attention to how prior distributions for the probability, location, and degree of violation are developed and used. Thus, Gilbert samples from plots that are next to plots already known to be in violation; the frequency of air sampling is tied to the frequency of past violations; and experienced plant inspectors come to know where the bodies may be buried and how to look for them.

Overall, this Conference was eminently successful in bringing out a broad range of problems, issues, and research needs. It has also provided some answers, though the most important products of our work here will continue to unfold for years to come. Our Chair, speakers, and discussants deserve much thanks for a job well done.

BIBLIOGRAPHY

1. Marcus AH. Time Scales: Biological, environmental, regulatory. This conference.
2. Hertzberg RC. Discussion of paper by Marcus. This conference.
3. Nelson WC. Statistical issues in human exposure monitoring. This conference.
4. Hunt WF. Discussion of paper by Nelson. This conference.
5. Bisgaard S, Hunter WG. Designing environmental regulations. This conference.
6. Johnson WB. Discussion of paper by Bisgaard and Hunter. This conference.
7. Price B. Quality control issues in testing compliance with a regulatory standard: Controlling statistical decision error rates. This conference.
8. Flatman GT. Discussion of paper by Price. This conference.
9. Gilbert RO, Miller ML, Meyer HR. On the design of a sampling plan to verify compliance with EPA standards foe radium-226 in soil at uranium mill tailings remedial action sites. This conference.
10. Chesson J. Discussion of paper by Gilbert, Miller, and Meyer. This conference.
11. Holley JW, Nussbaum BD. Distributed compliance: EPA and the lead bubble. This conference.
12. Ross NP. Discussion pf paper by Holley and Nussbaum. This conference.
13. Frank NH, Curran TC. Variable sampling schedules to determine PM$_{10}$ status. This conference.
14. Warren J. Discussion of paper by Frank and Curran. This conference.
15. Hammerstrom TS, Wyzga RE. Analysis of the relationship between maximum and average in SO$_2$ time series. This conference.
16. Bailey RC. Discussion of paper by Hammerstrom and Wyzga. This conference.
17. Jones CA. Models of Regulatory enforcement and compliance, with an application to the OSHA Asbestos Standard. Harvard University Economics Department, Unpublished doctoral dissertation, 1982.
18. Little RJA, Rubin DB. Statistical Analysis with Missing Data. John Wiley, 1987.
19. Jernigan RW. A Primer on Kriging. Statistical Policy Branch, US Environmental Protection Agency, 1986.

APPENDIX A: Program

Monday, October 5

INTRODUCTION

9:00 a.m. *Paul I. Feder*, Conference Chairman, Battelle Columbus Division
Dorothy G. Wellington, U.S. Environmental Protection Agency

I. TOXICOKINETIC AND PERSONAL EXPOSURE CONSIDERATIONS IN THE DESIGN AND EVALUATION OF MONITORING PROGRAMS

9:10 a.m. Time Scales: Biological, Environmental, Regulatory
Allan H. Marcus, Battelle Columbus Division
DISCUSSION
Richard C. Hertzberg, U.S. EPA, ECAO–Cincinnati

10:15 a.m. BREAK

10:30 a.m. Some Statistical Issues in Human Exposure Monitoring
William C. Nelson, U.S. EPA, EMSL–Research Triangle Park
DISCUSSION
William F. Hunt, Jr., U.S. EPA, OAQPS–Research Triangle Park

12:00 noon LUNCHEON

II. STATISTICAL DECISION AND QUALITY CONTROL CONCEPTS IN DESIGNING ENVIRONMENTAL STANDARDS AND COMPLIANCE MONITORING PROGRAMS

1:00 p.m. Designing Environmental Regulations
Soren Bisgaard, University of Wisconsin–Madison
DISCUSSION
W. Barnes Johnson, U.S. EPA, OPPE–Washington, D.C.

2:15 p.m. BREAK

2:30 p.m. Quality Control Issues in Testing Compliance with a Regulatory Standard: Controlling Statistical Decision Error Rates
Bertram Price, Price Associates, Inc.
DISCUSSION
George T. Flatman, U.S. EPA, EMSL–Las Vegas

III. COMPLIANCE WITH RADIATION STANDARDS

3:40 p.m. On the Design of a Sampling Plan to Verify Compliance with EPA Standards for Radium-226 in Soil at Uranium Mill Tailings Remedial Action Sites
Richard O. Gilbert, Battelle Pacific Northwest Laboratories; *Mark L. Miller*, Roy F. Weston, Inc.; *H.R. Meyer*, Chem-Nuclear, Inc.
DISCUSSION
Jean Chesson, Price Associates, Inc.

5:00 p.m. RECEPTION

(see over)

Tuesday, October 6

IV. THE BUBBLE CONCEPT APPROACH TO COMPLIANCE

9:00 a.m.	Distributed Compliance—EPA and the Lead Bubble *John W. Holley, Barry D. Nussbaum*, U.S. EPA, OMS–Washington, D.C. DISCUSSION *N. Philip Ross*, U.S. EPA, OPPE–Washington, D.C.
10:15 a.m.	BREAK

V. COMPLIANCE WITH AIR QUALITY STANDARDS

10:30 a.m.	Variable Sampling Schedules to Determine PM_{10} Status *Neil H. Frank, Thomas C. Curran*, U.S. EPA, OAQPS–Research Triangle Park DISCUSSION *John Warren*, U.S. EPA, OPPE–Washington, D.C.
12:00 noon	LUNCHEON
1:00 p.m.	The Relationship Between Peak and Longer Term Exposures to Air Pollution *Ronald E. Wyzga*, Electric Power Research Institute, *Thomas S. Hammerstrom, H. Daniel Roth*, Roth Associates DISCUSSION *R. Clifton Bailey*, U.S. EPA, OWRS–Washington, D.C.
2:15 p.m.	BREAK

SUMMARY OF CONFERENCE

2:30 p.m.	*John C. Bailar III*, McGill University, Department of Epidemiology and Biostatistics

This Conference is the final in a series of research conferences on interpretation of environmental data organized by the American Statistical Association and supported by a cooperative agreement between ASA and the Office of Standards and Regulations, under the Assistant Administrator for Policy Planning and Evaluation, U.S. Environmental Protection Agency.

Conference Chairman and Organizer:
Paul I. Feder, Battelle Columbus Division

APPENDIX B: Conference Participants

Ruth Allen
U.S. EPA
401 M Street, S.W., RD-680
Washington, DC 20460

Stewart J. Anderson
CIBA-GEIGY Corporation
556 Morris Avenue, SIC 249
Summit, NJ 07901

John C. Bailar III
(McGill University)
468 N Street, S.W.
Washington, DC 20024

R. Clifton Bailey
Environmental Protection Agency
401 M Street, S.W., WH-586
Washington, DC 20460

T. O. Berner
Battelle Columbus Division
2030 M Street, N.W., Suite 700
Washington, DC 20036

Soren Bisgaard
University of Wisconsin
Center for Quality & Productivity Improvement
Warf Building, 610 Walnut Street
Madison, WI 53705

Jill Braden
Westat, Inc.
1650 Research Boulevard
Rockville, MD 20852

Chao Chen
U.S. EPA
401 M Street, S.W., RD-689
Washington, DC 20460

Jean Chesson
Price Associates, Inc.
2100 M Street, N.W., Suite 400
Washington, DC 20037

James M. Daley
(U.S. EPA)
12206 Jennel Drive
Bristow, VA 22012

Susan Dillman
U.S. EPA
401 M Street, S.W., TS-798
Washington, DC 20460

Paul I. Feder
Battelle Columbus Division
505 King Avenue
Columbus, OH 43201

George T. Flatman
U.S. EPA, EMSL-LV
P.O. Box 93478
Las Vegas, NV 89193-3478

Paul Flyer
Westat, Inc.
1650 Research Boulevard
Rockville, MD 20852

Ruth E. Foster
U.S. EPA-OPPE/OSR
401 M Street, S.W.
Washington, DC 20460

Neil H. Frank
U.S. EPA, OAQPS
MD-14
Research Triangle Park, NC 27711

Richard O. Gilbert
Battelle Pacific Northwest Lab
P.O. Box 999
Richland, WA 99352

J. Hatfield
Battelle Columbus Division
2030 M Street, N.W., Suite 700
Washington, DC 20036

Richard C. Hertzberg
U.S. EPA, ECAO
Cincinatti, OH 45268

John W. Holley
(U.S. EPA-OMS)
9700 Water Oak Drive
Fairfax, VA 22031

William F. Hunt, Jr.
U.S. EPA, OAQPS
MD-14
Research Triangle Park, NC 27711

Thomas Jacob
Viar and Company
209 Madison
Alexandria, VA 22314

Robert Jernigan
American University
Department of Mathematics
and Statistics
Washington, DC 20016

W. Barnes Johnson
U.S. EPA, OPPE
401 M Street, S.W., PM-223
Washington, DC 20460

Herbert Lacayo
U.S. EPA
401 M Street, S.W., PM-223
Washington, DC 20460

Emanuel Landau
American Public Health Association
1015 15th Street, N.W.
Washington, DC 20005

Darlene M. Looney
CIBA-GEIGY Corporation
556 Morris Avenue, SIC 257
Summit, NJ 07901

Allan H. Marcus
Battelle Columbus Division
P.O. Box 13758
Research Triangle Park, NC 27709-2297

Lisa E. Moore
U.S. EPA
26 W. Martin Luther King Jr. Drive
Cincinnati, OH 45268

William C. Nelson
U.S. EPA, EMSL
MD-56
Research Triangle Park, NC 27711

Barry D. Nussbaum
U.S. EPA
401 M Street, S.W., EN-397F
Washington, DC 20460

Harold J. Petrimoulx
Environmental Resources
Management, Inc.
999 West Chester Pike
West Chester, PA 19382

Bertram Price
Price Associates, Inc.
2100 M Street, N.W., Suite 400
Washington, DC 20037

Dan Reinhart
U.S. EPA
401 M Street, S.W., TS-798
Washington, DC 20460

Alan C. Rogers
U.S. EPA
401 M Street, S.W.
Washington, DC 20460

John Rogers
Westat, Inc.
1650 Research Boulevard
Rockville, MD 20852

N. Philip Ross
U.S. EPA, OPPE
401 M Street, S.W., PM-223
Washington, DC 20460

Brad Schultz
U.S. EPA
401 M Street, S.W., TS-798
Washington, DC 20460

John Schwemberger
U.S. EPA
401 M Street, S.W., TS-798
Washington, DC 20460

Paul G. Wakim
American Petroleum Institute
1220 L Street, N.W.
Washington, DC 20005

John Warren
U.S. EPA, OPPE
401 M Street, S.W., PM-223
Washington, DC 20460

Dorothy G. Wellington
U.S. EPA
401 M Street, S.W., PM-223
Washington, DC 20460

Herbert L. Wiser
U.S. EPA
401 M Street, S.W., ANR-443
Washington, DC 20460

Ronald W. Wyzga
Electric Power Research Institute
P.O. Box 10412
Palo Alto, CA 94303

Conference Coordinator:
Mary Esther Barnes
American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3402